Chapter

1

Introduction

- Brief history of computer technology and its impact on society
- Purpose of the book and what readers can expect to learn
- Overview of the main themes and topics that will be covered

1.1 Brief History of Computer Technology and Its Impact on Society

Computer technology has its origin in the early 19th century when scientists and mathematicians started thinking of automating calculations. The earliest computing devices include the abacus, the Jacquard loom and Charles Babbage's Analytical Engine which paved the way for the development of modern computers [57].

The turning point was in the mid-20th century with the invention of the electronic computer. In 1946, the ENIAC (Electronic Numerical Integrator and Computer) became the world's first general-purpose electronic computer. It was quite large and needed a lot of power to work. Over the following decades, computer technology underwent substantial progress. The invention of transistors in the late 1940s replaced the use of bulky vacuum tubes in computers which made them smaller, faster and more reliable. The development of integrated circuits in the 1960s and 1970s led to the invention of microprocessors which transformed the field of computing by making it possible to build smaller, more powerful and affordable computers.

Personal computers (PCs) became available in the 1970s and 1980s, giving people the opportunity to have computing power at their disposal. The use of graphical user interfaces (GUIs) by companies such as Apple and Microsoft in the 1980s made computers easy to use and therefore available to a wider audience. The 1990s saw the rapid expansion of the internet which linked computers and allowed the sharing of information across the globe. This development ushered in a new age of communication, commerce, and collaboration.

Computer technology has made tremendous progress in the twenty-first century. The availability of smartphones, tablets, and wearable devices has made computing even more portable and more integrated into our daily lives. The advancement of cloud computing has made it possible to store and process big data which has in turn led to the development of artificial intelligence, machine learning and big data analytics [211, 173, 61].

Computer technology has had a significant effect on society. It has transformed different industries which include communication, education, healthcare, finance, entertainment, and transportation. Computers have enhanced the speed of information exchange, increased productivity and changed the way we work, learn and interact with others. However, alongside its benefits, computer technology has also brought some challenges. Some of the issues that have emerged as a result of computer technology include privacy, security, digital divide and ethical implications of artificial intelligence.

In conclusion, the history of computer technology has witnessed tremendous evolution from mechanical devices to modern systems that are interconnected. Its impact on society has been transformative, affecting how

we live, work, and communicate. As we move forward, managing the implications of computer technology will remain an important challenge.

1.2 Purpose of the Book and What Readers Can Expect to Learn

The main objective of this book is to present a detailed and fair analysis of the various consequences of rapid progress in computer technology on society since the last few decades.

Readers can expect the book to follow the main developments in computing and digital connectivity which have fundamentally altered our living patterns and working methods and learning methods and social interactions. The book will show the substantial advantages these technologies have brought as well as the dangers and disputes and difficulties which have arisen.

The book will address how artificial intelligence and automation transform work systems as well as the effects of big data analytics and social media platforms on free speech and democracy and privacy rights. The book intends to examine the critical questions about computer technology regulation and ethical development as well as social good distribution.

The book provides detailed analysis of the changes computerization has caused in law medicine commerce transportation finance and education without taking an overly negative or overly optimistic perspective. The aim of this book is to deliver readers an objective assessment of the trade-offs and various stakeholder perspectives. The book uses multiple fields like history and economics and law and ethics to create an interdisciplinary framework for understanding computer technology's social effects. The book presents an integrated understanding of this complex subject. Readers can anticipate a holistic view of this intricate subject.

This book highlights the current importance of its content since it examines both the present-day opportunities and challenges of modern technological developments. Through its analysis readers can understand both current events and emerging trends. Readers can expect the book to reveal effective methods for obtaining maximum benefits from new technologies while protecting public interest values including privacy, safety, accountability and fairness. This book will supply readers with proper context along with critical thinking abilities to assess computer technology impacts across time through balanced and accurate perspectives.

1.3 Overview of the Main Themes and Topics that will be Covered

The upcoming chapters introduce their core subjects through which readers can discover the complex patterns in the way technology affects society.

Chapter 2: Local and Global Impacts of Computing Technology on Work

This chapter investigates how computer technology has transformed work environments along with its local and global consequences. We discuss the historical development of automation as well as its employment effects which have been both beneficial and detrimental. We examine the rise of the gig economy and remote work alongside changes in job structures because of automation.

Chapter 3: Freedom of Speech in Cyberspace

The third chapter leads us into the modern concept of free speech in the digital world. We trace the historical development of this fundamental right and its application to the internet. The discussion examines both how social media and online platforms affect free speech and the challenges of regulating online content moderation while analyzing the government's regulatory powers in online discourse.

Chapter 4: Privacy Risks & Principles

This chapter investigates the essential element of personal privacy that defines our everyday existence. The research examines privacy violations and data breaches within the digital age by presenting specific examples

from the real world. The study explores both how surveillance affects personal privacy and how individuals can protect their information when dealing with the expanding digital environment.

Chapter 5: Crime & Security

The fifth chapter explores the dangerous aspects of the digital world through the lens of cybercrime. We analyze the different types of cybercrime together with their social implications. The chapter explores both the preventive measures and detection techniques for these crimes and their ethical aspects in cybersecurity and law enforcement activities to stop digital crimes.

Chapter 6: Intellectual Property

The sixth chapter examines how intellectual property law interacts with computer technology. The discussion covers digital copyright infringement together with patent infringement and trademark violations while also analyzing open-source software and the sharing economy's effects on intellectual property rights. The research examines the government's position regarding intellectual property protection.

Chapter 7: Evaluating & Controlling Technology

The chapter demonstrates why ethical evaluation remains crucial for technology development processes. We establish methods to evaluate societal technology impacts which consider different stakeholder viewpoints. A thorough evaluation of modern technological advancements such as AI and biotechnology for ethical concerns takes place in this chapter.

Chapter 8: Professional Ethics & Responsibilities

The last chapter investigates the moral duties that experts within the technology sector must fulfill. We present examples of ethical misconduct together with their social consequences. Professional organizations play a crucial role in establishing ethical conduct through their activities and we investigate their methods for maintaining ethical standards in modern technology development.

Our ultimate purpose is to foster informed discussions and promote the responsible use of technology in our rapidly changing world. We anticipate that after completing this educational journey you will gain knowledge about computer technology's effects on society while possessing the skills to help define its future direction. Our intellectual exploration starts now.

Chapter

2

Local and Global Impacts on Work

- The history of automation and the impact of computers on jobs
- Benefits and Risks of Using Technology in the Workplace: The Key to Success in Balancing Them
- The gig economy and the rise of remote work
- The future of work in the age of automation
- Technology's Dual Edges: Employee Communication and Employer Monitoring
- Navigating the Convergence: Separating or Merging Work and Personal Systems

2.1 The History of Automation and the Impact of Computers on Jobs: Evolution and Challenges

The evolution of automation, together with the influence of computers on jobs, represents a compelling timeline that modified industrial operations, modified professional responsibilities and opened fresh possibilities. The session examines how automation developed starting from its original stage to the computer era as well as its profound effects on job market development between positive and negative aspects [24, 25, 3, 5, 6, 44, 35, 34, 53, 37, 164, 217, 227, 125, 197, 189, 215, 51, 187].

The initial automation systems emerged from water clocks and ancient automata that existed in historical civilizations. Mechanized manufacturing brought the first substantial advancement in automation during the 18th century. Production processes underwent revolutionary changes through the combination of steam power and the assembly line which resulted in the mass production system that defined the twentieth century. The world experienced a fundamental transformation when computer technology emerged. The introduction of computers brought about a new dimension of precise automation with increased operational efficiency and advanced system complexity.

During the late twentieth century the microprocessor invention brought personal computing and robotics to the world after the digital era began. Digital automation through the analog to digital transition opened new possibilities for automation tasks that moved past mechanical operations to include mental processing. The automation landscape has not stopped its development since the beginning of the 21st century due to artificial intelligence (AI), machine learning, and the Internet of Things (IoT) which define the current smart automation period.

Impact on Job Roles: Computers introduced fundamental changes to the functions people performed in different professional domains. The job market underwent substantial changes because of automation which brings higher productivity and efficiency but causes major changes to employment.

1. Job Displacement: Computers together with automation have automated repetitive manual and routine tasks which can be easily automated. The takeover of data entry positions along with assembly line positions and basic calculation duties by machines caused employment losses for individuals who worked in these sectors.

2. Job Transformation: The implementation of automated systems resulted in substantial modifications to established job responsibilities. The takeover of repetitive tasks by machines enables humans to dedicate their time to higher-level responsibilities that require critical thinking, problem-solving, creativity and emotional intelligence. Automation has established new professional possibilities which enable workers to perform work that is both mentally demanding and emotionally rewarding.
3. Job Creation: The automation revolution has resulted in both job elimination and the establishment of new positions and industries. Modern technological advancements have given birth to new job sectors including software development, data analysis, cybersecurity, and artificial intelligence. These new positions need specialized abilities and knowledge to create positions available to individuals who learn and develop appropriate skills.
4. Skill Requirements: Automation has increased the demand for higher-level skills, such as problem-solving, critical thinking, creativity, and adaptability. Jobs which need complex decision-making alongside human interaction and creative thinking have shown lower susceptibility to automation.

Future Outlook: Technology will continue to advance thus continuing its influence on employment opportunities. The future job market will either have completely automated positions or require human-machine collaborative work. People along with communities need to learn new abilities and pursue lifelong education while focusing on work domains where human capabilities enhance automated systems.

Challenges and Considerations: The numerous advantages of automation exist alongside various obstacles and issues that need attention.

1. Job Polarization: The job market has become polarized through automation since low-skilled workers now stand at a distance from high-skilled workers. Routine work functions are most susceptible to automation which leads to the reduction of middle-skilled occupations. The current skill gap between automation workers and others requires continuous skill development for people to compete in an automated environment.
2. Job Insecurity and Transition: Automation produces job insecurity for people whose occupations become vulnerable to automation. The negative effects of automation on employment can be reduced through transitional support that includes retraining programs combined with educational initiatives and support networks for displaced workers.
3. Ethical Considerations: The growing use of automation systems requires ethical analysis to consider responsible implementation practices. The implementation of decision-making roles by computers and artificial intelligence systems leads to accountability and privacy concerns alongside bias and transparency issues. The responsible use of automation technologies needs ethical frameworks together with regulations to maintain proper standards.

In conclusion, the job market has experienced substantial change due to automation throughout its historical development from its first appearance to computer technology integration. Automation has resulted in the elimination of certain positions while it also generates fresh employment opportunities and improves production efficiency. To adapt to automation changes individuals and societies must adopt lifelong learning alongside skill development and active strategies to navigate the shifting work environment. Automation's potential can be maximized through responsible addressing of its challenges and ethical considerations to establish a future collaboration between humans and machines for innovative societal development.

2.2 Benefits and Risks of Using Technology in the Workplace: The Key to Success in Balancing Them

The workplace has undergone a transformation in its operations due to technology and employees' collaboration has also changed. Although technology has many benefits, it also has several disadvantages. This session discusses the benefits and disadvantages of using technology in the workplace and the need to strike a balance between the two in order to harness the benefits and minimize the negative consequences [25, 3, 6, 44, 34, 53, 37, 164, 217, 227, 125].

2.2.1 Advantages of Using Technology in the Workplace:

The implementation of technology in the workplace is one of the greatest changes that has occurred in business today. With the growing dependency of the global economy on digital infrastructure, the use of technology in the workplace is no longer a choice but a requirement. We now looks at the numerous advantages of the application of technology in the work environment, including increasing productivity, encouraging innovation, enhancing communication and promoting growth.

Enhanced Productivity and Efficiency: The primary advantage of technology in the workplace is the vast improvement in productivity and efficiency it brings to the workplace. Automation tools are able to complete tasks that are repetitive, thereby enabling human workers to concentrate on tasks that are more complex and creative. Software solutions simplify the workflow, reduce mistakes, and shorten the duration of tasks that used to take hours or even days to accomplish. For example, project management tools like Asana or Trello allow team members to track the progress of tasks in real time, set deadlines, and assign tasks with ease, thereby ensuring that projects are completed on time.

Improved Communication and Collaboration: Technology has greatly improved how employees interact and work together. With the introduction of communication platforms such as Slack, Zoom, and Microsoft Teams, team members can be in touch with each other even when they are not in the same location. This has made it possible for businesses to adopt the culture of remote work and work with global teams, which has made it possible to reach out to a wider talent pool. Also, cloud-based services like Google Workspace and Office 365 allow real-time collaboration on documents and projects, breaking the limits of location and time.

Data-Driven Decision Making: The use of technology in the workplace has brought about data-driven cultures where decisions are based on facts rather than feelings. Business intelligence tools and advanced analytics help companies gain knowledge on market trends, customer behavior and internal operations and make decisions accordingly. This is because businesses can use big data to improve their strategies in order to better serve consumers and outcompete their competitors.

Enhanced Security: Given the large amount of data that is processed daily, security is a major concern for businesses. Modern technology offers advanced security solutions including encryption, multi-factor authentication and cybersecurity software to prevent unauthorized access and cyber threats to sensitive information. This not only helps in the protection of the company's data but also helps in meeting the requirements of the law and gaining the trust of the customers.

Cost Savings: Even though the initial costs of investing in technology are high, the long-term costs are quite low. Digital tools eliminate the need for physical materials, reduce travel costs through virtual meetings, and minimize the need for physical office space through remote working capabilities. Also, technology can automate tasks that would otherwise require additional personnel, thus resulting in a more streamlined and cost-efficient workforce.

Employee Satisfaction and Talent Attraction: Technology in the workplace can improve employee satisfaction as employees are provided with tools that make their work easier and more enjoyable. For example, gamification elements can be added to the training programs to make learning more fun and improve retention. Companies that embrace cutting-edge technologies are often regarded as industry leaders, and this attracts top talent who want to work in an innovative environment.

Fostering Innovation: Technology is a catalyst for innovation. It allows employees to think differently and come up with new ideas on how to solve problems. Whether it is through advanced research and development tools, virtual reality simulations, or just by encouraging a digital creative culture, technology is a crucial aspect of the innovation process. The workplace benefits from numerous significant advantages through technological implementation. The advantages of workplace technology implementation include improved operational efficiency along with communication facilitation and decision-making support and security enhancement. The benefits of cost reduction together with satisfied employees and innovation promotion establish a competitive dynamic business environment. Digital transformation demands organizations to implement technology solutions in their workplaces because it enables future success as well as present-day performance maintenance. The investment produces an agile and robust organizational culture that becomes forward-thinking while enabling endless growth opportunities.

2.2.2 Disadvantages of Using Technology in the Workplace:

Despite the numerous benefits that technology has brought to the workplace, there are also a number of drawbacks. From the impact on human relationships to the potential for data breaches, the risks of a technologically saturated work environment can be serious for businesses and employees. This session now outlines the main disadvantages of relying heavily on technology in the workplace, and offers a balanced view of the digital revolution in the world of work.

Over-Reliance on Technology: The most significant disadvantage is the danger of over-reliance on technology. As tasks and processes become more automated or digitalized, there is a risk that employees may lose basic skills or the ability to do things manually. This dependency can be problematic when technology fails or during cyber-attacks when systems can be compromised, and employees cannot perform basic functions or solve problems on their own.

Privacy Concerns and Data Security: The integration of technology into the workplace has raised significant concerns over privacy. Surveillance tools used to monitor employee productivity can infringe on personal privacy and lead to a culture of mistrust. Furthermore, the collection of data, although useful for business intelligence, increases the risk of data breaches. Despite the latest security measures, the threat of cyber-attacks still persists, and the consequences of data leaks can be very costly in terms of reputation and financial stability for a company.

Increased Costs: Contrary to the cost-saving potential, technology can also introduce substantial costs. The initial outlay for state-of-the-art hardware and software can be quite expensive and there is the on-going expense of maintenance, updates, and training. For small businesses, these costs can be prohibitive and may result in an increase in the competitive gap between small and large enterprises.

Job Displacement: Automation and AI have the power to displace jobs, especially those jobs that involve doing repetitive tasks. While new jobs are created in the tech sector, they often require skills that the existing workforce may not possess. This displacement can lead to unemployment or underemployment for certain segments of the workforce, and this can contribute to economic inequality and social unrest.

Distraction and Reduced Human Interaction: Technology can be a source of constant distraction. The prevalence of email, instant messaging, and social media can fragment attention and reduce productivity. Moreover, the ease of digital communication can also diminish face-to-face interaction, which is very important for the relationships and the development of a team culture. The subtleties of human communication are often lost in digital exchanges, which can lead to misunderstandings and conflict.

Impacts on Health: The pervasive use of technology can have detrimental effects on physical and mental health. Prolonged screen time can lead to eye strain, poor posture, and repetitive strain injuries. Furthermore, the blurring of work-life boundaries, especially with mobile technology, can lead to stress, burnout, and difficulties in disengaging from work-related activities outside of regular working hours.

Resistance to Change: Not all employees are comfortable with the rapid pace of technological change. Learning new systems and adapting to new processes can be stressful and time consuming, especially for those who are less tech-savvy. This can lead to resistance, reduced morale, and even turnover, as employees may feel overwhelmed or undervalued.

While technology certainly enhances certain aspects of workplace functionality, it is not without its drawbacks. The disadvantages—from over-reliance and privacy concerns to health implications and resistance to change—require careful consideration and management. It is vital for organizations to strike a balance, ensuring that technology serves to support and not subvert the human elements of the workplace. Employers should foster an environment where technology is used responsibly and ethically, with adequate support and training for employees, and where personal connections and well-being are prioritized alongside productivity and innovation. In navigating the digital age, it is the harmonious integration of technology with human-centric values that will ultimately determine the success and health of the workplace.

In conclusion, the workplace has experienced a revolutionary transformation because of technology which brings multiple benefits to organizations and their employees. Organizations need to understand and solve the potential disadvantages along with challenges which technology brings to the workplace. Organisations must achieve an optimal technology implementation that maximizes efficiency and productivity but also protects against security threats and information overload as well as job loss risks.

The advantages and disadvantages of workplace technology implementation depend on the organization type and specific industry context as well as the particular workplace situation. Organizations need to evaluate both benefits and challenges before establishing security protocols and employee training programs and striking a balance between technological use and human-centered operations.

2.3 The Gig Economy and the Rise of Remote Work: Redefining the Modern Workforce

The rise of remote work coupled with the gig economy has revolutionized traditional work methods during recent years. The modern labor market operates through short-term contracts which accompany freelance work in the gig economy structure and remote work enables professionals to work from non-traditional locations. The analysis explores how the gig economy has evolved together with remote work while studying their effects on people as well as organizations and general society [13, 132, 193, 18, 38, 109, 116, 160, 181, 225, 228].

The Gig Economy: Technology advancements combined with changing work patterns have driven the fast-growing popularity of the gig economy. People now have the alternative choice to work independently by

taking either temporary projects or offering freelance services instead of seeking traditional employment. This trend offers several advantages. People benefit from enhanced flexibility through the system since they get to determine their work schedules along with the locations and methods. The gig economy presents job seekers with expanded work opportunities that enable skill diversification and professional experience growth. Through its platform the economy allows people to become independent business owners who can execute their personal ambitions.

The rise of the gig economy brings multiple difficulties to the workplace. Workers in the gig economy experience unstable income because their earnings depend on the availability of job opportunities. Many individuals who work as giggers receive minimal benefits and security since they do not qualify for medical coverage or retirement funds and face no job protection. Government officials along with businesses need to solve these problems to develop a sustainable and fair model for the gig economy which delivers benefits to both employees and employers.

The Rise of Remote Work: Technological improvements particularly high-speed internet collaborative tools and communication platforms have driven remote work to grow rapidly. The shift brings major effects for both business entities and individual staff members. Remote work gives employees the chance to balance their work and personal lives more effectively and work at flexible times. Employees benefit from this arrangement since it eliminates travel time and decreases expenses while allowing them to design their perfect workspace. Remote work enables people with physical disabilities and those living in distant locations to enter the workforce through its inclusive nature.

Remote work enables businesses to recruit talent from all around the world thus expanding their talent pool beyond geographical limitations. Companies gain access to top professionals across the world through remote work which leads to better diversity and greater innovation. Business entities that adopt flexible work models experience lower office expenses while achieving higher productivity levels because they match work environments to employee preferences. The adoption of remote work results in environmental sustainability because it minimizes the carbon emissions that stem from office infrastructure and commuting needs.

Nevertheless, remote work poses challenges. Teamwork suffers because physical distance makes communication and collaboration more difficult so organizations must develop efficient virtual tools and methods. Remote work can negatively affect mental health because of isolation and social isolation which requires organizations to build social connection spaces and support systems for remote employees. Remote work creates obstacles in maintaining proper work-life separation because work and home environments blend together so self-discipline together with defined boundaries should be implemented.

In conclusion, the modern workforce undergoes fundamental change because of the growing popularity of remote work and the gig economy which redefine established employment structures. These systems deliver workers autonomy while offering flexibility and opportunities to build their own careers. Businesses gain multiple advantages from the gig economy by acquiring larger talent pools while achieving higher productivity levels at reduced costs.

The ongoing challenges of income variability together with benefits restrictions and work-life balance maintenance need to be addressed.

The transformation of work requires policymakers and businesses and workers to build adaptable systems along with protective measures which follow work's changing nature. Through proper management of remote work benefits and gig economy challenges society can develop an inclusive and agile sustainable work environment for the future.

2.4 The Future of Work in the Age of Automation: Navigating Opportunities and Challenges

The modern world experiences rapid technological changes through artificial intelligence robotics and machine learning which transform how work operates. Understanding the upcoming changes to work environments requires investigation into automation effects on the future job landscape. This session investigates the benefits and obstacles of modern times because humans need to adapt and learn new skills and reshape their human role in automated environments.

Job Transformation through Automation: The implementation of automation technologies holds the potential to transform both industries and job functions. Machines excel at executing repetitive and routine operations which enables human employees to perform creative and critical thinking tasks that also require emotional intelligence. The work transformation provides people with the chance to pursue tasks that bring meaning and intellectual engagement to their careers.

Automation brings about job displacement as a major concern despite its benefits. Machines will replace specific job duties which may lead to the elimination of particular job roles. The existence of automation should not be mistaken for job elimination because it leads to job transformations instead. The evolving nature of technology requires employees to learn how to work effectively with machines so they can use technological abilities to enhance their individual abilities.

Upskilling and Lifelong Learning: The automation era demonstrates why people need to constantly enhance their abilities through lifelong learning and upskilling. Rapid technological advancements require individuals to learn through a growth mindset while actively seeking opportunities to build new skills. Digital literacy and data analysis and problem-solving and creative thinking will become essential skills for the job market because of this development.

Organizations together with governments must create learning opportunities that are both accessible and inclusive. Educational institutions must team up with businesses and policymakers to develop skills that match the changing needs of the modern workforce. This approach enables workers to move from automated positions into roles that demand human-specific capabilities.

The Changing Nature of Work: Automation technologies will transform how work exists in the world. The traditional employment model featuring permanent full-time jobs may develop into adaptable working patterns. The gig economy together with freelance work and remote work options will expand in popularity to enable people to start entrepreneurial ventures while achieving better work-life equilibrium. The successful adoption of these changes requires new social safety net systems which should include fair and inclusive work protections.

Collaboration between Humans and Machines: Humans and machines will work together to create the future of work. Machines demonstrate exceptional capabilities for handling large datasets and pattern detection and executing repetitive operations with high accuracy. The strengths of human workers include emotional intelligence together with creative thinking and complex problem-solving and adaptability. The combination of human innovation with machine computational strength allows us to produce unmatched productivity levels and societal advancement together with innovative breakthroughs.

In conclusion, the upcoming age of automation brings about a revolutionary and energetic work environment of the future. People will find purposeful work through intellectual stimulation if they learn to adapt their skills. The partnership between humans and machines will lead to groundbreaking productivity achievements

and innovative breakthroughs. The transition requires proactive steps to resolve employment disruptions and education deficits and work environment transformation.

The ability to learn continuously combined with inclusive skill development programs and acceptance of work evolution will help us handle automation through resilience while technology drives human progress. Strategic planning and human-focused collaboration will lead the age of automation to create a future which delivers shared prosperity and fulfillment.

2.5 Technology's Dual Edges: Employee Communication and Employer Monitoring

The introduction of technology transformed workplace communication by enabling employees to communicate through various instant methods as well as between employees and their employers. The improved communication methods enabled by employers introduce new monitoring possibilities that create challenges regarding workplace privacy and employee trust. The examination discusses how technology affects workplace communication together with the monitoring methods employers use and the necessary equilibrium between operational success and staff privacy rights [15, 28, 93, 126, 216].

Impact on Employee Communication: The workplace has experienced a significant transformation through technological advancements in its communication methods. Modern workplaces rely more on emails together with instant messaging applications and collaborative platforms than they did on traditional memos and in-person meetings for quick real-time information exchange. The technology tools have removed geographical limitations which led to the development of remote work and global team collaboration that prove essential in today's interconnected mobile society.

Video conferencing technologies provide an inclusive communication method which combines visual cues with presence features to improve remote team members' mutual understanding and connection. Through social intranet platforms and enterprise social networks organizations enable employees to exchange both official information and foster social connections which strengthen their workplace community.

Implications of Employer Monitoring: These same technological tools have enabled employers to access previously unreachable levels of employee activity monitoring. Modern workplaces implement different technologies which enable them to monitor both employee email activities and their typing actions as well as their internet usage and physical movements. The main purpose of workplace monitoring includes security needs and company policy compliance together with protecting valuable information and maximizing worker efficiency.

The extensive monitoring methods which companies implement produce negative consequences for their operations. The combination of workplace monitoring generates workplace distrust which lowers staff morale and may violate personal employee privacy rights. The psychological effects of continuous monitoring tend to generate higher stress levels while making employees feel undervalued and mistrusted by their organization. The surveillance practices known as "Big Brother" suppress employee creativity and initiative because monitored employees avoid risk-taking and thinking outside their roles.

Legal and Ethical Considerations: The laws that protect employee privacy show great complexity in different jurisdictions throughout the world. Organizations need to understand many privacy rules across different regions while multinational companies face additional challenges in complying with these regulations. Ethically, the practice of monitoring also presents dilemmas. Employers have a right to protect their business and optimize productivity but must respect their employees' dignity and self-determination. The ethical application of monitoring technology requires businesses to provide transparency about their monitoring practices and to establish clear policies which maximize employee privacy protection.

Balancing Communication Advantages with Monitoring Concerns: Organizations need to find a proper equilibrium between their technology-based communication systems and employee privacy rights. The implementation of monitoring needs careful consideration because it requires both staff participation and their approval for the process. Clear policies about monitoring activities and their purposes help to reduce employee concerns and establish necessary boundaries. Staff members can gain effective understanding of communication technology through training which promotes responsible use.

In conclusion, workplace technology delivers substantial advantages in efficiency and teamwork but simultaneously poses risks to privacy and trust. Both employers and employees must maintain an open dialogue about digital tool usage in the professional environment. Organizations should implement ethical monitoring practices together with transparency and respect-based policies to benefit from technological advancements without infringing on employee rights and well-being. The relationship between technology and employees should develop as a synergy that supports employee capabilities while preserving their independence.

2.6 Navigating the Convergence: Separating or Merging Work and Personal Systems

The integration of work and personal systems into one system has become more common in modern times. The advancement of mobile technology along with the growing trend of remote work and the disappearance of professional boundaries has sparked a discussion about how to handle digital tools and information. This session studies the effects of both work-personal system separation and integration while presenting the advantages and disadvantages for each method.

Separating Work and Personal Systems: The traditional model of professional and personal life boundaries applies to digital systems just as it does to other aspects. The separation requires distinct devices together with separate accounts and applications for work tasks and personal activities.

Advantages:

1. Security and Privacy: When work information exists on separate systems from personal data it minimizes potential security threats. The system protects both workplace privacy from supervisory oversight as well as personal privacy.
2. Clarity and Organization: Work-specific tools create organized digital environments which prevent the confusion caused by mixing personal and professional data.
3. Work-Life Balance: A distinct separation between work and personal systems enables people to switch between their roles effectively which results in better work-life separation.
4. Professional Image: Work systems which operate independently from personal systems help organizations project professionalism because they minimize the chance of personal content accidentally appearing in professional contexts.

Challenges:

1. Increased Costs: Operating dual system structures demands considerable financial investment to acquire additional devices and software products.
2. Inefficiency: Moving between different systems creates difficulties which results in workflow interruptions and efficiency problems.
3. Repetitive work involves creating duplicate copies of contacts calendars and files between different systems which leads to time-consuming efforts and possible inconsistencies.

Merging Work and Personal Systems: Merging work and personal systems takes place through integrating professional and personal activities on shared platforms and devices which happens because of flexible work culture requirements.

Advantages:

1. One system which handles work and personal activities gives users maximum convenience because all tools and information remain accessible from this single location.
2. Merging systems creates financial advantages because organizations no longer need to purchase duplicate hardware or software solutions.
3. Integrated systems enable flexible work patterns that match the changing requirements for employee mobility and flexibility in current work environments.

Challenges:

1. Security Concerns: The security protocols of corporate environments may not apply to personal devices and networks which could put sensitive information at risk.
2. Distractions: Work performance suffers when personal notifications interrupt the professional environment.
3. Privacy Intrusions: Workers must sometimes grant their employers access to work devices which violates their private rights.
4. Boundary Erosion: Work invasions into personal time result in stress and burnout because employees lack defined boundaries.

In conclusion, organizations must decide between separate and combined work and personal systems based on individual circumstances. A thorough evaluation of workplace culture and personal needs as well as work nature and individual preferences becomes necessary. The decision to separate work and personal activities requires a detailed understanding of all involved trade-offs.

Workers who need to protect sensitive information or maintain complete work-life separation will choose separate systems. Those who value both work flexibility and minimalism often choose combined systems yet implement advanced security protocols and disciplined self-management practices.

The goal is to achieve maximum productivity and security while preserving personal boundaries and well-being through proper balancing. The ongoing technological evolution will drive new approaches to handle professional-personal digital life boundaries which requires sustained human-centered digital choice evaluation.

2.7 Case Study 1

RESEARCH ARTICLE

Automation and the changing nature of work

Cecily Josten, Grace Lordan*

Department of Psychological and Behavioural Science, The London School of Economics and Political Science, London, United Kingdom

* *g.lordan@lse.ac.uk*

Abstract

This study identifies the job attributes, and in particular skills and abilities, which predict the likelihood a job is recently automatable drawing on the Josten and Lordan (2020) classification of automatability, EU labour force survey data and a machine learning regression approach. We find that skills and abilities which relate to non-linear abstract thinking are those that are the safest from automation. We also find that jobs that require 'people' engagement interacted with 'brains' are also less likely to be automated. The skills that are required for these jobs include soft skills. Finally, we find that jobs that require physically making objects or physicality more generally are most likely to be automated unless they involve interaction with 'brains' and/or 'people'.

1. Introduction

Research on the automation and the future of work has brought with it a range of research contributions, which seek to determine which occupations will be lost to automation. For example, Frey and Osborne [1] estimate the susceptibility of occupations to computerization and find that 47% of US occupations are at risk of automation, and point to service jobs as being susceptible to automation. Many other contributions in the automation literature rely on defining automatable work through measures of the tasks associated with a particular occupation rather than the occupation overall to get a more nuanced understanding of the impact of automation on employment [2–4]. One of the most prominent is owed to Autor and Dorn [5] and Autor et al. [6] who define a job as automatable if it is high in routine task-intensity. Specifically, routine task-intensity is defined based on how high a job ranks on routine content, and how low it ranks on abstract and manual content. Information on the routine, abstract and manual task content of each respective occupation comes from the US *Dictionary of Occupation Titles* where incumbents are asked to grade their occupation with respect to particular attributes. A job is then defined as automatable if it is in the top third of the distribution of routine task-intensity. This measure of automatable work has followed the big movements in the occupation distribution accurately over the last decades–namely the hollowing out of the middle of the occupational distribution [7]. To this end the types of occupations available have become more polarized, with the majority of occupations falling into high and low skill categories, and mid skill jobs disappearing in numbers [8].

Further, Arntz et al. [2] estimate the risk of automation for 21 OECD countries also using a task-based approach. They argue that other studies using an occupation rather than a task-based approach such as Frey and Osborne [1] overestimate the risk of automation, partly due to the fact that cross-country variation is not taken into account other than through differences in occupation structure. They highlight that technical possibilities of automation do not equate to actual automation because that may be hindered by legal or ethical obstacles. They find that overall only 9% of jobs are automatable. Similarly, Nedelkoska and Quintini

https://doi.org/10.1371/journal.pone.0266326 Used with CC 4.0 license

[9] also study automation of jobs in OECD countries using the Survey of Adult Skills (PIAAC). They focus on very detailed occupational categories and the tasks therein and find that 14% of jobs in OECD countries are at high risk of being automated and 32% have a probability of being automated of 50% to 70%. Gregory et al. [4] emphasize that while routine occupations have been replaced between 1999 and 2010, product demand has also increased as a result of routine-replacing technological change leading to a net employment growth. Also focusing on routine-task intensity, Lewandowski et al. [3] focus on cross-country differences in the level of routineness in occupations. They find lower levels of routineness in high-skill occupations in countries with a higher GDP per capita. They also find correlations between computer use, higher education and higher literacy skills in a given country with routine task intensity.

While much of the automation literature relies on past employment data, the rapid progress on robotics, artificial intelligence (AI) and automation technologies has also motivated predicting automation developments in the near future [10, 11]. The importance of this exercise is bellied in Lordan [12] and Lordan and Neumark [13] who suggest that new jobs are now being automated; particularly jobs traditionally at the bottom of the occupation distribution. Further, advances in AI and in particular machine learning will likely affect at least some tasks in most occupations and will hence also disrupt jobs at the top of the income distribution [14].

Concretely, Webb [8] studies the impact of automation on occupational tasks and matches information on job tasks to patents issued for robots, software and AI to identify which tasks can be automated by different technologies to derive an exposure to automation score. He uses Google patents data, the O*NET database of occupations and tasks and US census data. O*NET is a database of occupations and tasks published by the US Department of Labor that provides detailed descriptions of a large number of occupations and has been used frequently in the literature studying the impact of automation and technical innovation on employment [5, 6, 8, 15]. He first analyzes the impact of his 'exposure to automation score on employment using historical data on robots and software patents and job descriptions and then repeats this exercise using patents on AI to predict future employment effects. AI is studied with respect to future developments as it is a relatively new phenomenon as compared to software and robotic innovations. While innovation on robots and software has mainly affected low skill and low wage occupations in the past, he finds that AI is increasingly predicted to disrupt high-skill occupations. Building on this work, Tolan et al. [16] link research intensity in AI to abilities required for specific job tasks using European survey data, O*NET data and AI benchmarking platforms. They find that jobs that were originally classified as non-automatable are increasingly affected by automation such as medical doctors. They find that abilities particularly affected by automation are abilities for idea creation while people abilities are less affected.

Lordan and Josten's [10] study is also forward-looking and takes the occupations classified by Autor and Dorn (2013) as given while reclassifying the remaining occupations as automatable depending on the number of patents recently available for each specific occupation thereby also predicting which jobs will be automated in the near future. They try to capture the most recent wave of automation by using patent developments in artificial intelligence, robots and automation more broadly as a proxy for technology that will be on the market shortly. If for any given occupation the authors find a large number of patents and find that successful patent pilots have been covered by the media, this occupation is classified as being on track to become automatable. A full list of the occupations classified as automatable by Lordan and Josten [10] can be found in S1 Table. This study builds on and uses the classification by Lordan and Josten [10] to analyze which job attributes and requirements predict the likelihood that a job is reclassified as automatable under their new definition. It thereby speaks to the literature on the automatability of tasks, skills and abilities.

In particular we use the automatability indicator by Lordan and Josten [10] matched with employment data from the European Labor Force Survey (EU-LFS) and with data on the skills and abilities required on the job from O*NET. The EU-LFS covers employment statistics of households from EU member countries, Switzerland, Sweden and the UK quarterly. We then regress the automatability indicator on the skills

and abilities respectively to analyze which skills and which abilities required in different occupations are susceptible to automation. This is then further linked to work by Lordan and Pischke [17] who capture the 'people', 'brains' and 'brawn' content of occupations with different risks of being automated; i.e. the extent to which an occupation involves people interaction, cognitive thinking skills or physicality respectively.

Identifying future job requirements by occupations is crucial to be able to quantify changes in demand for skills and abilities resulting from automation [18]. First, this information is relevant to policymakers and companies interested in the future of work. In particular, it informs on the skills and abilities required of workers in the near future as well as the activities they will likely be engaged in. Second, these findings further help conversations surrounding the re-organization of education and other development activities to ensure the stock and flow of the relevant skills for the 4th industrial revolution. The returns to education are constantly increasing with the rise in technological progress with specific skills such as digital and non-cognitive skills becoming particularly important [19]. Third, this information also helps to gain a more nuanced understanding of the exact aspects of the occupations at risk of automation and hence addresses criticism by Arntz et al. [20], among others, who find that focusing on the automatability of occupations rather than job tasks overstates the risk of automation and omits important aspects of the automation developments.

2. Data

Job Level Abilities and Skills Proxies

To capture the abilities and skills required by occupations at the three-digit occupation level we draw on O*NET version 15. O*NET is an occupational database by the US Department of Labor that narrowly defines occupations with respect to the tasks and activities and the skills and abilities required on the job. This database has been used frequently in the automation literature both in the US and the UK [2, 3, 5, 16]. The difference in task content of occupations in the US versus the UK has been shown to be small, further justifying using this resource to classify occupation attributes using UK data [3, 21].

Specifically, O*NET offers 80 distinct items in the abilities classification. The first column of Table 2 lists each of these items and column (2) provides a brief description of the item. The third column of Table 2 documents the secondary category a specific ability is in, while column (4) provides an overall category. In addition to abilities, O*NET offers 40 distinct items in the skills classification. Again the first column of Table 3 lists each of these items and column (2) provides a brief description each item. The third column documents the secondary category a specific ability is in, while column (4) provides an overall category. Given the large number of abilities (80) and skills (40) items, the secondary and overall category in columns (3) and (4) from Tables 2 and 3 will be utilized respectively for the interpretation of the analysis.

EU LFS and Lordan and Josten

Our analysis relies on data from Lordan and Josten [10] who match data from the European Labour Force Survey (EU-LFS) between 2013–2016 to their automation classification. The EU-LFS is conducted across all Member States of the European Union, Iceland, Norway, Switzerland and the United Kingdom and consists of quarterly collected household data on employment. We include all countries that have data in the years we analyze, in addition to 3-digit occupation codes. A full list of the countries we include can be found in Table 2.1.

Lordan and Josten [10] match their definition of recently automatable jobs based on patents at the 3-digit code occupation level to the EU-LFS using crosswalks provided by Lordan and Pischke [17]. To derive the automation classification, Lordan and Josten [10] revise 216 occupations that have been classified as non-automatable by Autor and Dorn [6] and search for the occupation name together with either the term "robot,"

"automation," or "artificial intelligence" in Google patents and then also in Google News. Depending on the number of patents and/or news articles, an occupation is then classified as either fully automatable, polarized automatable (i.e. partially automatable) or non-automatable.

Table 2.1 shows the shares of automatable employment (i.e. polarized and fully automatable) by EU-LFS country. We first note that the Lordan and Josten [10] classifications suggest that a large share of jobs in every country is recently automatable. Specifically, Finland is the country that has the lowest share of jobs that are classified as recently automatable (approximately 21%). For the remaining Scandinavian countries (Norway, Sweden and Denmark), Table 2.1 suggests that about 37% of the jobs are automatable, similar to the shares for the Netherlands and Iceland. In contrast, for Croatia, Czech Republic, Italy and Latvia more than 50% of the jobs are recently automatable. Overall, Table 2.1 highlights that there is large variation in the share of occupations that can be automated across EU-LFS countries.

Table 2.1 Lordan and Josten (2020) shares of employment in 2013–2016 by country.

	Recently Automatable
Austria	0.516
Belgium	0.443
Croatia	0.565
Cyprus	0.512
Czech Republic	0.516
Denmark	0.370
Estonia	0.458
Finland	0.207
France	0.478
Germany	0.474
Greece	0.574
Hungary	0.544
Iceland	0.374
Ireland	0.481
Italy	0.508
Latvia	0.525
Lithuania	0.527
Luxembourg	0.442
Netherlands	0.374
Norway	0.363
Portugal	0.485
Spain	0.474
Sweden	0.378
Slovakia	0.513
UK	0.405

https://doi.org/10.1371/journal.pone.0266326.t001

3. Methodology

The main analysis in this work relies on EU-LFS data. Our main analysis relies on an indicator variable that is equal to 1 if an occupation is newly automatable under the Lordan and Josten [10] classification and 0 otherwise. Jobs that are denoted as automatable by Author and Dorn [5] are excluded from the analysis given that these jobs were classified as automatable based on O*NET occupation attributes (i.e. routine, manual and abstract tasks), so they mechanically relate to the O*NET attributes. In addition, this exclusion allows us to clarify the differences in skills and abilities between non-automatable jobs and newly automatable jobs. If these jobs were included this delineation would not be possible.

We proceed by regressing the indicator variable on each set of job attributes and skills respectively as provided by O*NET. We control for differences across country with a set of country fixed effects and for differences across time with a set of year fixed effects. We have two main sets of regressions. The first set regresses the automation indicator variable on the 80 ability domains and the second regresses it on the 40 skills domains. We apply Lasso regression analysis, a shrinkage and a variable selection method for linear regression models. This approach is chosen as we wish to reduce the dimensionality of the abilities and skills variables under consideration. The goal of a Lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The Lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero. Variables with a regression coefficient equal to zero after the shrinkage process are excluded from the model. That is, these variables do not explain variation in the propensity for a job to be recently automatable and will be shown with a value of zero in the results tables. The remaining variables with a positive sign are those that describe the core skills and abilities that are most likely to become redundant because of the most recent wave of automation. In contrast, the remaining variables with a negative sign describe the core skills and abilities that are most likely to become more valuable. All non-zero variables are significant at the 1% significance level.

When estimating the Lasso for the abilities attributes we include 60 abilities attributes, in addition to country and year fixed effects. Country fixed effects control for location specific factors that are time invariant. Given the small time window we do not expect that time varying factors will distort our results.

4. Results

Abilities

The results from the pooled country analysis for abilities are documented in Table 2.2 We first note that the regression considering abilities explains 97% of the variation in the Lordan and Josten [10] automation indicator. That is, we can explain almost all of the variation in the indicator variable with these measures of ability. Second, if we take column (3) and column (4) there is no single secondary or overall ability category that persistently has negative or positive signs. Rather, within each of these categories there are abilities that are becoming less important and others that are becoming more important. For example, arm-hand steadiness is a fine manipulative psychomotor ability (e.g. the ability to keep your hand and arm steady while moving your arm) that is highly unlikely to be automated given the estimates. In contrast, finger dexterity which is also a fine manipulative psychomotor ability (i.e. the ability to make precisely coordinated movements of the fingers of one or both hands to grasp, manipulate, or assemble very small objects) is likely to be automated. Similarly, fluency of ideas is an idea generation and reasoning ability which relates to the ability to come up with a number of ideas about a topic (i.e. the number of ideas is important, not their quality, correctness, or creativity) which is highly unlikely to be automated. In contrast, problem sensitivity (i.e. recognizing that there is a problem but not solving the problem) in the same category is highly likely to be automated.

https://doi.org/10.1371/journal.pone.0266326

Table 2.2 Abilities estimates.

(1)	(2)	(3)	(4)	(5)
ONET Item	**Description**	**Secondary category**	**Overall category**	**Coefficient**
Arm-Hand Steadiness	The ability to keep your hand and arm steady while moving your arm or while holding your arm and hand in one position.	Fine manipulative abilities	Psychomotor abilities	-3.587
Auditory Attention	The ability to focus on a single source of sound in the presence of other distracting sounds.	Auditory and speech abilities	Sensory Abilities	0.000
Category Flexibility	The ability to generate or use different sets of rules for combining or grouping things in different ways.	Idea Generation and Reasoning Abilities	Cognitive Abilities	-2.381
Control Precision	The ability to quickly and repeatedly adjust the controls of a machine or a vehicle to exact positions.	Control Movement Abilities	Psychomotor Abilities	1.663
Deductive Reasoning	The ability to apply general rules to specific problems to produce answers that make sense.	Idea Generation and Reasoning Abilities	Cognitive Abilities	-4.904
Depth Perception	The ability to judge which of several objects is closer or farther away from you, or to judge the distance between you and an object.	Visual Abilities	Sensory Abilities	0.558
Dynamic Flexibility	The ability to quickly and repeatedly bend, stretch, twist, or reach out with your body, arms, and/or legs.	Flexibility, Balance and Coordination	Physical Abilities	2.784
Dynamic Strength	The ability to exert muscle force repeatedly or continuously over time. This involves muscular endurance and resistance to muscle fatigue.	Physical Strength	Physical Abilities	-0.256
{2,3Explosive Strength	The ability to use short bursts of muscle force to propel oneself (as in jumping or sprinting), or to throw an object.	Physical Strength	Physical Abilities	0.555
Extent Flexibility	The ability to bend, stretch, twist, or reach with your body, arms, and/or legs.	Flexibility, balance and coordination	Physical Abilities	-2.347
Far Vision	The ability to see details at a distance.	Visual Abilities	Sensory Abilities	-2.219

https://doi.org/10.1371/journal.pone.0266326

(1) ONET Item	(2) Description	(3) Secondary category	(4) Overall category	(5) Coefficient
Finger Dexterity	The ability to make precisely coordinated movements of the fingers of one or both hands to grasp, manipulate, or assemble very small objects.	Fine manipulative abilities	Psychomotor abilities	2.169
Flexibility of Closure	The ability to identify or detect a known pattern (a figure, object, word, or sound) that is hidden in other distracting material.	Perceptual Abilities	Cognitive abilities	0.000
Fluency of ideas	The ability to come up with a number of ideas about a topic (the number of ideas is important, not their quality, correctness, or creativity).	Ideas generation and reasoning abilities	Cognitive abilities	-2.830
Glare sensitivity	The ability to see objects in the presence of glare or bright lighting.	Visual abilities	Sensory abilities	-3.635
Gross Body Co-ordination	The ability to coordinate the movement of your arms, legs, and torso together when the whole body is in motion.	Flexibility, Balance and Coordination	Physical Abilities	0.266
Gross Body Equilibrium	The ability to keep or regain your body balance or stay upright when in an unstable position.	Flexibility, Balance and Coordination	Physical Abilities	0.000
Hearing Sensitivity	The ability to detect or tell the differences between sounds that vary in pitch and loudness.	Auditory and Speech abilities	Sensory Abilities	-1.139
Inductive Reasoning	The ability to combine pieces of information to form general rules or conclusions (includes finding a relationship among seemingly unrelated events).	Idea Generation and Reasoning Abilities	Cognitive Abilities	0.000
Information Ordering	The ability to arrange things or actions in a certain order or pattern according to a specific rule or set of rules (e.g., patterns of numbers, letters, words, pictures, mathematical operations).	Idea Generation and Reasoning Abilities	Cognitive Abilities	0.000

(*Continued*)

Table 2.2 Abilities estimates. (*Continued*)

(1)	(2)	(3)	(4)	(5)
ONET Item	**Description**	**Secondary category**	**Overall category**	**Coefficient**
Manual Dexterity	The ability to quickly move your hand, your hand together with your arm, or your two hands to grasp, manipulate, or assemble objects.	Fine Manipulative Abilities	Psychomotor Abilities	0.000
Math Reasoning	The ability to choose the right mathematical methods or formulas to solve a problem.	Quantitative Abilities	Cognitive Abilities	1.253
Memorization	Abilities related to the recall of available information.	Quantitative Abilities	Cognitive Abilities	0.000
Multi Limb Co-ordination	The ability to coordinate two or more limbs (for example, two arms, two legs, or one leg and one arm) while sitting, standing, or lying down. It does not involve performing the activities while the whole body is in motion.	Control movement abilities	Psychomotor Abilities	1.604
Near Vision	The ability to see details at close range (within a few feet of the observer).	Near Vision	Visual Abilities	0.000
Night Vision	The ability to see under low light conditions.	Near Vision	Visual Abilities	3.025
Number Facility	The ability to add, subtract, multiply, or divide quickly and correctly.	Quantitative Abilities	Cognitive Abilities	-1.817
Oral Comprehension	The ability to listen to and understand information and ideas presented through spoken words and sentences.	Verbal Abilities	Cognitive Abilities	0.000
Oral Expression	The ability to communicate information and ideas in speaking so others will understand.	Verbal Abilities	Cognitive Abilities	4.378
Originality	The ability to come up with unusual or clever ideas about a given topic or situation, or to develop creative ways to solve a problem.	Idea generation and Reasoning Abilities	Cognitive Abilities	0.000

https://doi.org/10.1371/journal.pone.0266326

(1) ONET Item	(2) Description	(3) Secondary category	(4) Overall category	(5) Coefficient
Perceptual Speed	The ability to quickly and accurately compare similarities and differences among sets of letters, numbers, objects, pictures, or patterns. The things to be compared may be presented at the same time or one after the other. This ability also includes comparing a presented object with a remembered object.	Perceptual Abilities	Cognitive Abilities	-4.204
Perceptual Vision	The ability to see objects or movement of objects to one's side when the eyes are looking ahead.	Visual Abilities	Sensory Abilities	0.713
Problem Sensitivity	The ability to tell when something is wrong or is likely to go wrong. It does not involve solving the problem, only recognizing there is a problem.	Idea generation and Reasoning Abilities	Cognitive Abilities	2.190
Rate Control	The ability to time your movements or the movement of a piece of equipment in anticipation of changes in the speed and/or direction of a moving object or scene.	Control Movement Abilities	Psychomotor Abilities	3.112
Reaction Time	The ability to quickly respond (with the hand, finger, or foot) to a signal (sound, light, picture) when it appears.	Reaction Time and Speed Abilities	Psychomotor Abilities	2.509
Response Orientation	The ability to choose quickly between two or more movements in response to two or more different signals (lights, sounds, pictures). It includes the speed with which the correct response is started with the hand, foot, or other body part.	Control Movement Abilities	Psychomotor Abilities	-4.984
Selective Attention	The ability to concentrate on a task over a period of time without being distracted.	Attentiveness	Cognitive Abilities	-0.963
Sound Localisation	The ability to tell the direction from which a sound originated.	Auditory and Speech Abilities	Sensory Abilities	-1.685

(*Continued*)

https://doi.org/10.1371/journal.pone.0266326

Table 2.2 Abilities estimates. (*Continued*)

(1)	(2)	(3)	(4)	(5)
ONET Item	**Description**	**Secondary category**	**Overall category**	**Coefficient**
Spatial Orientation	The ability to know your location in relation to the environment or to know where other objects are in relation to you.	Spatial Abilities	Cognitive Abilities	1.051
Speech Clarity	The ability to speak clearly so others can understand you.	Auditory and Speech Abilities	Sensory Abilities	0.000
Speech Recognition	The ability to identify and understand the speech of another person.	Auditory and Speech Abilities	Sensory Abilities	-2.099
Speed of Closure	The ability to quickly make sense of, combine, and organize information into meaningful patterns.	Perceptual Abilities	Cognitive Abilities	5.164
Speed of Limb Movement	The ability to quickly move the arms and legs.	Reaction Time and Speed Abilities	Psychomotor Abilities	-1.118
Stamina	The ability to exert yourself physically over long periods of time without getting winded or out of breath.	Endurance	Physical Abilities	0.557
Static Strength	The ability to exert maximum muscle force to lift, push, pull, or carry objects.	Physical Strength Abilities	Physical Abilities	0.679
Time Sharing	The ability to shift back and forth between two or more activities or sources of information (such as speech, sounds, touch, or other sources).	Attentiveness	Cognitive Abilities	2.796
Trunk Strength	The ability to use your abdominal and lower back muscles to support part of the body repeatedly or continuously over time without'giving out' or fatiguing.	Physical Strength Abilities	Physical Abilities	0.000
Visual Color Discrimination	The ability to match or detect differences between colors, including shades of color and brightness.	Visual Abilities	Sensory Abilities	4.322

https://doi.org/10.1371/journal.pone.0266326

(1) ONET Item	(2) Description	(3) Secondary category	(4) Overall category	(5) Coefficient
Visualisation	The ability to imagine how something will look after it is moved around or when its parts are moved or rearranged.	Spatial Abilities	Cognitive Abilities	-0.752
Wrist-Finger Speed	The ability to make fast, simple, repeated movements of the fingers, hands, and wrists.	Reaction Time and Speed Abilities	Psychomotor Abilities	-1.251
Written Comprehension	The ability to read and understand information and ideas presented in writing.	Verbal Abilities	Cognitive Abilities	0.000
Written Expression	The ability to communicate information and ideas in writing so others will understand.	Verbal Abilities	Cognitive Abilities	-0.295
R Squared = 0.97				2701297

https://doi.org/10.1371/journal.pone.0266326.t002

Although it may seem counter-intuitive that within the same ability category there are pairs of abilities that are both automatable and non-automatable, the description of each item (see column (2) Table 2) highlights a logical theme. In general, an ability is automatable if it entails an underlying action that is repeatable or follows some logic. For example, control precision relies on adjusting controls quickly and repeatedly. The repetitive nature of this ability implies it is easily codifiable and machines have a comparative advantage in speed of execution. In contrast, speech recognition, which involves understanding spoken language, is unlikely to be automated. This finding is intuitive as it is difficult to predict what someone will say unless it entails common conversations such as those in telephone customer care. For example, customer care in banking now frequently utilize artificial intelligence to direct calls or provide bank balances but it is human operators that deal with complaints and other issues.

The overall conclusion from studying Table 2 is that jobs that are high on 'brains' (i.e. involve abstract thinking) are far less likely to be automated. In this case, 'brains' is short-hand for thinking and can involve reacting to other individuals (e.g. in caring or teaching professions), performing a service (e.g. as mechanic or fine dining waiter) or engaging in agile or creative thinking (e.g. in a leadership or knowledge worker role). Occupations that are predicted to be automated are low on 'brains' and high in routine contents. The abilities analysis is hence largely in line with the work by Author and Dorn [6] who predict that jobs that involve routine tasks will be automated. However, given that in this analysis we only focus on the occupations classified as non-automatable by Author and Dorn [6] it further reiterates that automation has continued to make progress automating jobs that are high in routine content.

Skills

The results from the pooled country analysis for skills are documented in Table 3. This regression explains 84% of the variation in the Lordan and Josten [10] automation indicator. Analyzing columns (3) and (4) of Table 3, the results are consistent with the analysis of abilities in that there is again no secondary or overall skills category that has a persistent negative or positive effect on the automation indicator. In addition and also consistent with the abilities analysis, there is a pattern which suggests that 'brains' (i.e. thinking and reacting) is becoming more important while on the job as compared to routine work.

For example, within the overall category of basic skills (see column 4), the O*NET item active learning is a skill that is more likely to be automated. This is consistent with machines being able to process a large amount of information quickly. However, the O*NET items critical thinking and monitoring of performance, which essentially involve using information that is available to pass judgement and make decisions, are less likely to be automated. Within cross functional skills, skills that center around identifying potential problems, setting rules and gathering information are most likely to be automated (for example the O*NET items of time management, operation monitoring, system evaluation and management of human resources). In contrast, those skills that require manual actions (e.g. equipment maintenance), communicating information (for example instructing), leadership (e.g. persuasion) and knowledge work (e.g. systems analysis and technology design) are less likely to be automated.

Table 2.3 Skills estimates.

(1)	(2)	(3)	(4)	(5)
ONET Item	**Description**	**Secondary category**	**Overall category**	**Coefficient**
Active Learning	Understanding the implications of new information for both current and future problem-solving and decision-making.	Process	Basic Skills	1.630
Active Listening	Giving full attention to what other people are saying, taking time to understand the points being made, asking questions as appropriate, and not interrupting at inappropriate times.	Content	Basic Skills	0.608
Complex Problem Solving	Identifying complex problems and reviewing related information to develop and evaluate options and implement solutions.	Complex Problem Solving	Cross Functional Skills	-0.580
Coordination	Adjusting actions in relation to others' actions.	Social Skills	Cross Functional Skills	-6.025
Critical Thinking	Using logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions or approaches to problems.	Process	Basic Skills	-3.233
Equipment Maintenance	Performing routine maintenance on equipment and determining when and what kind of maintenance is needed.	Technical Skills	Cross Functional Skills	-0.504
Equipment Selection	Determining the kind of tools and equipment needed to do a job.	Technical Skills	Cross Functional Skills	0.000
Installation	Installing equipment, machines, wiring, or programs to meet specifications.	Technical Skills	Cross Functional Skills	0.540

https://doi.org/10.1371/journal.pone.0266326

(1)	(2)	(3)	(4)	(5)
Instructing	Teaching others how to do something.	Social Skills	Cross Functional Skills	-2.752
Judgement and Decision Making	Considering the relative costs and benefits of potential actions to choose the most appropriate one.	Systems Skills	Cross Functional Skills	2.775
Learning Strategies	Selecting and using training/ instructional methods and procedures appropriate for the situation when learning or teaching new things.	Systems Skills	Cross Functional Skills	0.000
Management of Financial Resources	Determining how money will be spent to get the work done, and accounting for these expenditures.	Resource Management Skills	Cross functional Skills	0.000
Management of Material Resources	Obtaining and seeing to the appropriate use of equipment, facilities, and materials needed to do certain work.	Resource Management Skills	Cross functional Skills	-1.262
Management of Personnel Resources	Motivating, developing, and directing people as they work, identifying the best people for the job.	Resource Management Skills	Cross functional Skills	2.264
Mathematics	Using mathematics to solve problems.	Content	Basic Skills	0.000
Monitoring	Monitoring/Assessing performance of yourself, other individuals, or organizations to make improvements or take corrective action.	Process	Basic Skills	-4.676
Negotiation	Bringing others together and trying to reconcile differences.	Social Skills	Cross functional Skills	1.967
Operation Monitoring	Watching gauges, dials, or other indicators to make sure a machine is working properly.	Technical Skills	Cross Functional Skills	1.831
Operation and Control	Controlling operations of equipment or systems.	Technical Skills	Cross Functional Skills	0.000
Operations Analysis	Analyzing needs and product requirements to create a design.	Technical Skills	Cross Functional Skills	-1.091
Persuasion	Persuading others to change their minds or behavior.	Social Skills	Cross functional Skills	-0.827

(*Continued*)

https://doi.org/10.1371/journal.pone.0266326

Table 2.3 Skills estimates. (*Continued*)

(1)	(2)	(3)	(4)	(5)
Programming	Writing computer programs for various purposes.	Technical Skills	Cross functional Skills	0.000
Quality Control Analysis	Conducting tests and inspections of products, services, or processes to evaluate quality or performance.	Technical Skills	Cross functional Skills	0.000
Reading Comprehension	Understanding written sentences and paragraphs in work related documents.	Content	Basic Skills	-2.474
Repairing	Repairing machines or systems using the needed tools.	Technical Skills	Cross Functional Skills	-0.935
Science	Using scientific rules and methods to solve problems.	Content	Basic Skills	1.044
Service Orientation	Actively looking for ways to help people.	Social Skills	Cross Functional Skills	-0.854
Social Perceptiveness	Being aware of others' reactions and understanding why they react as they do.	Social Skills	Cross Functional Skills	0.000
Speaking	Talking to others to convey information effectively.	Content	Basic Skills	1.235
System Analysis	Determining how a system should work and how changes in conditions, operations, and the environment will affect outcomes.	Systems Skills	Cross Functional Skills	-2.530
System Evaluation	Identifying measures or indicators of system performance and the actions needed to improve or correct performance, relative to the goals of the system.	Systems Skills	Cross Functional Skills	4.365
Technology Design	Generating or adapting equipment and technology to serve user needs.	Technical Skills	Cross Functional Skills	-2.203
Time Management	Managing one's own time and the time of others.	Resource Management Skills	Cross Functional Skills	4.214
Troubleshooting	Determining causes of operating errors and deciding what to do about it.	Technical Skills	Cross Functional Skills	0.000

https://doi.org/10.1371/journal.pone.0266326

(1)	(2)	(3)	(4)	(5)
Writing	Communicating effectively in writing as appropriate for the needs of the audience.	Content	Basic Skills	1.686
R squared = 0.84			N =	2701297

https://doi.org/10.1371/journal.pone.0266326.t003

A second notable pattern emerges if we reconsider Table 3. In general, the skills that require active interactions with people (i.e. implying that there is at least a two-way dialogue where an employee is reacting to the other person(s)) are not automatable. In essence, these jobs are an interaction between 'brains' and 'people' skills. In Table 3, these include coordination, instructing, monitoring and persuasion. This observation seems to suggest that jobs that involve 'people' interacted with 'brains' skills are also less likely to be automated.

Conclusion from the Skills and Abilities Analysis

From the estimates documented in Tables 2 and 3 we make the following three conclusions:

1. We can explain almost all of the variation in the jobs that are newly defined as automatable by Lordan and Josten [10] using the O*NET items of abilities or skills.
2. Jobs that require 'brains' (i.e. abstract and non-linear thinking) are far less likely to be automated as compared to jobs which require linear and codifiable thinking skills and abilities. At the top of the income distribution, jobs that require non-linear thinking may need critical thinking, decision-making and creativity. Elsewhere in the income distribution these jobs require skills that have been traditionally delivered in apprenticeships, from mechanics and carpenters to florists and hairdressers.
3. Jobs that require 'people' engagement interacted with 'brains' are also less likely to be automated. These jobs include management across all levels, coordinators of all types, teachers, carer and medical practitioners (including nursing). The skills and abilities that are required for these jobs include soft skills. The value of these skills in terms of adult outcomes has become a topic of recent writings in economics (for example Heckman and Kautz (2013) [22]; Kautz et al., (2014) [23] and Lordan and McGuire (2019) [24]) and has been recently noted specifically as skills that will be needed in the advent of the Fourth Industrial Revolution [12, 25].

Activities

To summarize further the conclusions described in conclusions 2. and 3. from the skills and abilities analysis (i.e. that jobs which require 'brains' and 'people' interacted with 'brains' are those that are the least likely to be automated) we consider a third analysis on work activities and context of a person's occupation. This third analysis is a replication of Table 3 in Lordan and Josten [10]. We later go further than Lordan and Josten [10] and present these estimates separately for each country in our dataset to allow for cross country comparisons.

To replicate their work we first follow Lordan and Pischke [17] and create three factors that represent the 'people' 'brains' and 'brawn' content in each three-digit code occupation. That is:

1. 'People' is a variable, which distils the information from many domains in activities and context that relate to having interactions with people on a day-to-day basis.
2. 'Brains' is a variable which distils the information from many domains in activities and context that relate to abstract thinking.

3. 'Brawn' is a variable which distils the information from many domains in activities and context that relate to interacting physically with objects, including making them on a daily basis.

We then regress our dummy representing whether a job is automatable on the 'people', 'brains' and 'brawn' variables (using ordinary least squares (OLS)). Consistent with Lordan and Josten [10]) we also interact the three variables with each other, implying we also include in our regression people*brains, people*brawn and brains*brawn. We again control for fixed country differences and yearly differences in the regression. The results from this analysis are documented in Table 2.4.

Table 2.4 'People', 'brains' and 'brawn' estimates.

Variable	Marginal effect
People	0.009***
	(0.000)
Brains	-0.070***
	(0.000)
Brawn	0.032***
	(0.000)
People * Brains	-0.002***
	(0.000)
People* Brawn	0.003***
	(0.000)
Brains* Brawn	0.000***
	(0.000)
N	2,698,151
R squared	42%

Notes *, **, *** denotes significance at the 10%, 5% and 1% levels respectively.
https://doi.org/10.1371/journal.pone.0266326.t004

Turning to Table 4, we can explain 42% of the variation in our automatable employment indicator with the three latent factors. The size and significance of the negative coefficient on the 'brain' factor strongly implies that jobs which require thinking are those that are safe from the most recent wave of automation. In addition, the interaction between the 'people' and 'brains' factors is negative and significant. This highlights that jobs which require thinking and social interactions are safer in the most recent wave of automation as compared to those that simply involve people interactions (as evidenced by the positive and significant coefficient on the 'people' factor). We note that the estimates in Table 4 imply that jobs that are high on 'brawn' content are those that are most likely to be automated in the most recent wave of automation. These jobs include making objects and physically lifting items. Overall, these conclusions align well with those that came from the abilities and skills analyses in Tables 2 and 3 respectively as they highlight the importance of abstract thinking and the combination of abstract thinking and people skills.

Country-Level Analysis

The S1 Appendix contains the estimates from Table 2 through 4 by country including all EU-LFS countries as depicted in Table 1. The differences across these estimates is driven only by differences in the occupation distribution for each country, while the classification of whether a job is recently automatable is fixed across time and country. We note that the separate analyses of skills and abilities for each country allows us to draw

similar conclusions to the pooled country estimates in Tables 2 and 3. That is, the abilities and skills that are becoming more important relate to the ability and skill to use 'brains' for abstract, strategic and creative thinking and the ability and skill to use 'brains' interacted with 'people'.

The 'people', 'brains' and 'brawn' categories allow us to summarize the differences and commonalities across countries most succinctly. These are:

1. For all countries, jobs that are high on 'brains' are least likely to be automated in the most recent wave of automation.
2. For many countries the 'people' coefficient is also negative and significant, implying that jobs that are 'people' facing are relatively safe from automation, regardless of their interaction with 'brains'. These countries are: Austria, Belgium, Cyprus, Czech Republic, Denmark, Germany, Estonia, Spain, France, Ireland, the Netherlands, Norway, Sweden.
3. For almost all countries there is a negative and significant relationship between the potential for automation and the 'people' and 'brains' interaction. The exceptions are Cyprus, Estonia, Ireland, Norway, Sweden and the UK. We note that for Estonia, Ireland, Norway and Sweden both the 'people' and 'brains' effects are independently negative and significant, highlighting that 'people' jobs in general are unlikely to disappear significantly. For Cyprus and the UK, the estimates suggest that 'brains' are the most important skills and abilities to develop given the current distribution of jobs.
4. In general, the brawn effect is positive and significant, implying that jobs that are high on 'brawn' content are at risk from automation. The exception of countries where this effect is negative are Cyprus, Czech Republic and Estonia.
5. The interaction effect between 'people' and 'brawn' is in general positive, significant but small in magnitude for just over half the countries in our study. In contrast, it is negative and significant but small in magnitude for: Finland, France, Hungary, Ireland, Iceland, Latvia, Norway, Sweden, Slovakia and the UK. The difference seems to be driven by a relatively large number of jobs in these countries in the low end of the income distribution that require physical lifting and people (for example, cleaning and caring).
6. The interaction effect between 'brains' and 'brawn' is in general negative, significant but small in magnitude. There are only three exceptions. These are: Cyprus, Czech Republic and Ireland.

5. Conclusion

This study identifies the job attributes which predict the likelihood that a job is recently automatable. In particular, it looks at the i) abilities and ii) skills required on the job and how they link to automatability. A third analysis (iii) considers the 'people', 'brains' and 'brawn' content of an occupation, i.e. the extent to which an occupation involves people interactions, abstract thinking and physicality respectively. The three analyses are also done on a country level to see how the impact of automatability on the labour market differs across countries.

Overall, we find that skills and abilities which relate to non-linear abstract thinking, which we term 'brains', are those that are the safest from automation. We also find that jobs that require 'people' engagement interacted with 'brains' are also less likely to be automated. The skills and abilities that are required for these jobs include soft skills. Finally, we find that jobs that require physicality (e.g. creating objects manually) are most likely to be automated unless they involve interaction with 'brains' and/or 'people'.

https://doi.org/10.1371/journal.pone.0266326

These findings are in line with the literature on the growing importance of cognitive and social skills for the future of work. In particular, Deming [25] finds that the interaction between cognitive and social skills has seen greater wage and employment growth, which is comparable to our finding of the importance of 'brains' alongside 'people' skills and abilities. It also matches findings of studies that link skills endowments or demand for skills to labour market outcomes and find that social and cognitive skills are increasingly rewarded and that there is a complementarity across those two dimension [26, 27].

Information and knowledge on future job requirements by occupations and by country is essential when trying to predict the demand for skills and abilities and activities going forward. It is important knowledge for policymakers and companies who can adapt policies and organizational settings regarding the future of work accordingly and ensure that individuals are prepared for current developments and what is yet to come. In particular, it informs conversations surrounding the re-organization of education and other development activities to ensure that the stock and flow of skills are ready for the Fourth Industrial Revolution. The returns to education are constantly increasing with the rise in technological progress with specific skills such as digital and non-cognitive skills becoming particularly important [19]. And this information also helps to gain a more nuanced understanding of the exact aspects of the occupations at risk of automation rather than just predicting automation overall and hence extends previous work. While we summarize our findings at the 'people', 'brains' and 'brawn' level, we still show and have briefly discussed the results by each O*NET abilities and skills item, which is informative to the reader interested in specific aspects of occupations and their automatability.

The differences in effects found at the country level likely reflect the fact that the structure of jobs and skills within country differ, coupled with each country being on a different trajectory with respect to automation. In addition, the policies that can protect jobs from automation also differ within country. A better understanding of such within country policies, coupled with their interaction with the labour market is an area for future research.

Supporting Information

S1 Table. Occupations classified by Josten and Lordan (2020) as automatable.
(DOCX)

S1 Appendix. Individual country analysis.
(DOCX)

Author Contributions

Conceptualization: Cecily Josten, Grace Lordan.
Data curation: Grace Lordan.
Formal analysis: Grace Lordan.
Funding acquisition: Grace Lordan.
Investigation: Grace Lordan.
Methodology: Grace Lordan.
Project administration: Grace Lordan.
Resources: Grace Lordan.
Software: Grace Lordan.
Supervision: Grace Lordan.
Validation: Grace Lordan.
Visualization: Grace Lordan.
Writing – original draft: Grace Lordan.
Writing – review & editing: Cecily Josten, Grace Lordan.

https://doi.org/10.1371/journal.pone.0266326

References

1. Frey CB, Osborne MA. The future of employment: how susceptible are jobs to computerization? Technol Forecast Soc Change. 2013;1–72.
2. Arntz M, Gregory T, Zierahn U. The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis Working Papers No. 189 The Risk of Automation for Jobs in OECD Countries A COMPARATIVE ANALYSIS. OECD Soc Employment, Migr Work Pap. 2016;(May):0_1.
3. Lewandowski P, Park A, Hardy W, Yang D. Technology, Skills, and Globalization: Explaining International Differences in Routine and Nonroutine Work Using Survey Data. IZA Discuss Pap [Internet]. 2019 [cited 2021 Jul 9]; Available from: www.iza.org
4. Gregory T, Salomons A, Zierahn U. Racing With or Against the Machine? Evidence from Europe. IZA Discuss Pap. 2019;
5. Autor D, Dorn D. The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market. Am Econ Rev. 2013;103(5):1553–97.
6. Autor DH, Dorn D, Hanson GH. Untangling Trade and Technology: Evidence from Local Labour Markets. Econ J. 2015;125(584):621–46.
7. Acemoglu D, Restrepo P. Tasks, automation, and the rise in US wage inequality. NBER Work Pap [Internet]. 2021;No. 28920. Available from: http://www.nber.org/papers/w28920
8. Webb M. The Impact of Artificial Intelligence on the Labor Market. SSRN. 2019;
9. Nedelkoska L, Quintini G. Automation, skills use and training (OECD Working Paper). OECD Soc Employment, Migr Work Pap [Internet]. 2018;(202):1–125. Available from: https://doi.org/10.1787/2e2f4eea-en
10. Josten C, Lordan G. Robots at Work: Automatable and Non-automatable Jobs. In: Handbook of Labor, Human Resources and Population Economics. Springer International Publishing; 2020. p. 1–24.
11. Innovation Vivarelli M., Employment and Skills in Advanced and Developing Countries: A Survey of Economic Literature. J Econ Issues [Internet]. 2014 [cited 2019 Nov 14];48(1):123–54. Available from: https://www.tandfonline.com/action/journalInformation?journalCode=mjei20
12. Lordan G. Robots at work—A report on automatable and non-automatable employment shares in Europe. Eur Comm—Dir Employment, Soc Aff Incl. 2018;
13. Lordan G, Neumark D. People versus machines: The impact of minimum wages on automatable jobs. Labour Econ. 2018 Jun 1;52:40–53.
14. Brynjolfsson E, Mitchell T, Rock D. What Can Machines Learn and What Does It Mean for Occupations and the Economy? Am Econ Assoc. 2018;108:43–7.
15. Frey CB, Osborne MA. The future of employment: How susceptible are jobs to computerisation? Technol Forecast Soc Change. 2017 Jan 1;114:254–80.
16. Tolan S, Pesole A, Martínez-Plumed F, Fernández-Macías E, Hernández-Orallo J, Gómez E. Measuring the Occupational Impact of AI: Tasks, Cognitive Abilities and AI Benchmarks. J Artif Intell Res. 2021;71:191–236.
17. Lordan G, Pischke J-S. Does Rosie like riveting? Male and female occupational choices. NBER Work Pap [Internet]. 2016 [cited 2019 Nov 27]; Available from: http://www.nber.org/data-appendix/w22495

18. Djumalieva J, Lima A, Sleeman C. Classifying Occupations According to Their Skill Requirements in Job Advertisements. ESCoE Discuss Pap [Internet]. 2018 [cited 2021 Jul 5]; Available from: www.escoe.ac.uk.

19. Goos M, Arntz M, Zierahn U, Gregory T, Carretero Gómez S, González Vázquez I, et al. The impact of Technological innovation on the Future of Work. JRC Work Pap Ser Labour, Educ Technol [Internet]. 2019 [cited 2021 Jul 5]; Available from: https://ec.europa.eu/jrc

20. Arntz M, Gregory T, Zierahn U. Revisiting the risk of automation. Econ Lett. 2017 Oct 1;159:157–60.

21. Fernández-Macías E, Hurley J, Bisello M. What do Europeans do at work?: A task-based analysis: European Jobs Monitor 2016. Publ Off Eur Union [Internet]. 2016 [cited 2021 Jul 9]; Available from: www.eurofound.europa.eu

22. Heckman J, Kautz T. Fostering and Measuring Skills: Interventions that improve character and cognition. Nber Work Pap Ser. 2013;

23. Kautz T, Heckman JJ, Diris R, Weel B ter, Borghans L. Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success. Natl Bur Econ Res [Internet]. 2014 Dec 15 [cited 2021 Jul 12]; Available from: https://www.nber.org/papers/w20749

24. Lordan G, McGuire A. Widening the High School Curriculum to Include Soft Skill Training: Impacts on Health, Behaviour, Emotional Wellbeing and Occupational Aspirations. IZA Discuss Pap. 2019;(12439).

25. Deming DJ. The Growing Importance of Digital Skills in the Workplace. Q J Econ. 2017;132(October):1593–640.

26. Deming D, Kahn LB. Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals. J Labor Econ. 2018;36(S1).

27. Weinberger C. The Increasing Complementarity between Cognitive and Social Skills. Rev Econ Stat [Internet]. 2014 [cited 2021 Jan 20];96(5). Available from: http://www.mitpressjournals.org/doi/abs/10.1162/REST_a_00449

2.8 Case Study 2

RESEARCH ARTICLE

The global polarisation of remote work

Fabian Braesemann[1,2,3], Fabian Stephany[1,2,3], Ole Teutloff[3,4], Otto Kässi[1,5], Mark Graham[1], Vili Lehdonvirta[1]

1 Oxford Internet Institute, University of Oxford, Oxford, United Kingdom, 2 Humboldt Institute for Internet and Society, Berlin, Germany, 3 DWG Datenwissenschaftliche Gesellschaft Berlin, Berlin, Germany, 4 Copenhagen Center for Social Data Science, University of Copenhagen, Copenhagen, Denmark, 5 Etla Economic Research Helsinki, Helsinki, Finland

Abstract

The Covid-19 pandemic has led to the rise of digitally enabled remote work with consequences for the global division of labour. Remote work could connect labour markets, but it might also increase spatial polarisation. However, our understanding of the geographies of remote work is limited. Specifically, in how far could remote work connect employers and workers in different countries? Does it bring jobs to rural areas because of lower living costs, or does it concentrate in large cities? And how do skill requirements affect competition for employment and wages? We use data from a fully remote labour market—an online labour platform—to show that remote platform work is polarised along three dimensions. First, countries are globally divided: North American, European, and South Asian remote platform workers attract most jobs, while many Global South countries participate only marginally. Secondly, remote jobs are pulled to large cities; rural areas fall behind. Thirdly, remote work is polarised along the skill axis: workers with in-demand skills attract profitable jobs, while others face intense competition and obtain low wages. The findings suggest that agglomerative forces linked to the unequal spatial distribution of skills, human capital, and opportunities shape the global geography of remote work. These forces pull remote work to places with institutions that foster specialisation and complex economic activities, i. e. metropolitan areas focused on information and communication technologies. Locations without access to these enabling institutions—in many cases, rural areas—fall behind. To make remote work an effective tool for economic and rural development, it would need to be complemented by local skill-building, infrastructure investment, and labour market programmes.

Introduction

The Covid-19 pandemic has made remote work the 'new normal' form of organisation for many white collar employees. Pre-covid, only the most flexible employers allowed employees to work at a distance. The coordination costs of managing remote teams were considered too high [1]. In forcing office employees to work from home, the pandemic has vastly accelerated the adoption of digital technologies [2] and organisational adjustments that allow business processes to operate productively at a distance [3].

Organisations can realise substantial cost savings and tap into global talent pools if they adopt remote working practices and start to outsource business processes to the remote workforce [4]. But what are the consequences of remote working for the global division of labour? Here, we investigate the global geographies of a fully remote labour market—a so-called online labour platform—that provides the digital

https://doi.org/10.1371/journal.pone.0274630 Used with CC 4.0 license

infrastructure to hire remote workers on demand. These online platforms have been established over the past 10 to 15 years and they allow even small companies or individual employers to outsource knowledge work to individual freelancers [5]. Due to the digital organisation of the hiring and work process on the platform, online labour markets can be considered as a prototype of a fully remote labour market. Having started as niche marketplaces for IT freelancers, these platforms now cover the whole spectrum of knowledge work, from data entry to management consulting, with millions of platform workers involved globally [6–8], and rising adoption during the Covid-19 pandemic [9]. The platform labour market has in fact seen substantial growth in recent years. According to Kässi et al. (2021) the number of globally registered online workers was 163 million in 2021 [6]. With the decade-long shift towards remote work [10] that has only been accelerated by the pandemic, more of the overall labour market could begin to resemble the online labour market soon.

Remote work organised via online platforms could bring jobs to workers from all over the globe [11–13]. In doing so, emote work could help to mitigate the global imbalance between the increasing supply of, and competition between, highly educated graduates in Global South and Global North countries—a phenomenon which has been described as the 'Global Auction' [14]—and the high global demand for talent. In bringing jobs and income to people in countries across the world and, in particular, in rural areas, remote work could help to foster more resilient, sustainable local communities [15, 16] and offer alternatives to the physical migration to places with more jobs and higher wage levels [17], if platform work can provide sustainable sources of income. However, several studies have reported that remote platform work is shaped by geographical frictions and biases that restrict participation, for example, time-zone differences, language-based communication difficulties, domestic and ethnic connections, or information asymmetries [18–21]. Similar to other complex economic activities such as research, innovation and industry [22] remote platform work might cluster in large cities. In the global remote labour market, modularisation of tasks and competitive dynamics could cause uneven geographical participation rates [23], bad working conditions [24–26], and precariousness for workers [27–29]; a process that has been subsumed under the term 'Digital Taylorism' [30]. Overall, the individual contributions provide coherent, but fragmented perspectives on the phenomenon of remote platform work. The body of previous contributions in sum, however, does not provide a comprehensive and stringent picture with respect to the global geography of remote platform work and its' impact on the relationship between urban and rural areas. One reason for this is a lack of sufficiently granular data.

Here, we examine the global geography of remote freelance work mediated by online labour platforms empirically based on a data set covering 1.8 million remote jobs from 2013 to 2020 to show that remote platform work mirrors the geographical and skill-based polarisation of labour markets at large [31]. The data comes from one platform, which is among the largest in terms of global market shares [8, 32]. In the discussion section, we consider what implications our findings may have for remote platform work more generally and other types of remote work, including regular employment performed remotely over the Internet.

In mapping the platform jobs to sub-national geographies in 139 countries and an established occupation taxonomy, we reveal that the remote labour market is polarised along three dimensions: globally between countries, between urban and rural areas within countries, and, overall, between occupations and skill-sets. We relate the spatial and occupational variation to differences in the global distribution of infrastructure, economic opportunities, and human capital. The data suggest that agglomerative forces shape the global geography of remote work. These forces pull the most profitable remote jobs to metropolitan areas and locations with existing competitive advantages in information and communication technologies. At the same time, rural areas and disadvantaged regions, particularly in Global South countries, are not able to attract

https://doi.org/10.1371/journal.pone.0274630

many remote jobs. Urban specialists can realise a premium for scarce skills [33], while less specialised remote platform workers compete for low-wage jobs.

Based on these findings we argue that the global distribution of skills and enabling institutions are the focal point of remote work. In this perspective, Digital Taylorism—the standardisation and modularisation of complex production processes of the knowledge economy broken down into simple and codified tasks together with improved monitoring capabilities [27]—is the very process that makes remote work and global digital value chains possible [14]. Enabling cost savings and access to talent pools simultaneously, Digital Taylorism is considered as the driver of the specialisation and global integration of the digital workforce via remote work. This process affects incomes and opportunities of knowledge work globally.

According to this reading of the empirical findings, we provide the following interpretation of the global geography of platform work. Skill-biased technological change [34–36] allows people with advanced digital skills (e. g. Data Scientists) to realise a premium from increased demand, while offshoring, computerisation, and global competition for jobs that require less specialised knowledge (e. g. Data Entry) drive wages downwards [37, 38]. The result is a polarised global market for knowledge work [39] with its' geography stratified along the lines of the unequal distribution of human capital. The antagonism between urban and rural areas—described by Paul Collier in his book *The Future of Capitalism* as a deep dividing line between the 'booming metropolis' and the 'broken provincial city' [40]—plays out fully in the remote labour market. This is because the institutions that enable a successful participation—access to knowledge building, training and professional networks—concentrate in urban environments. Rural regions are less able to offer specialised work opportunities and urban lifestyle [41, 42]. In contrast, in metropolitan areas, a highly specialised local economy creates an abundance of opportunities to maintain a tech-savvy 'creative class' [43, 44]. The most profitable remote jobs require specialised IT-skills and go therefore to metropolitan areas. The polarising forces that pull remote jobs to centres of economic activity with existing competitive advantages and digital infrastructures work almost unrestricted in the platform economy. This is because there are only limited frictions of geographical boundaries, labour market regulations or formal entry barriers, which increase the role of information asymmetries, uncertainties, trust cues, and reputation systems in online labour markets [45–47].

In summary, our main argument is that the unequal global distribution of remote work is the result of the unbalanced distribution of skills, human capital, and opportunities across the globe [48, 49]. This uneven distribution of economic conditions and competitive advantages transcends to the platform economy and drives the geographical polarisation of the remote labour market. Remote workers with access to hard-to-copy skills can realise a substantial premium, while those who lack marketable, digital skills participate in a global rat race for remote jobs. Digitally enabled remote work will likely not disperse knowledge work more evenly across space but rather reinforce prevalent agglomerative dynamics.

Literature Review

In the following, we review the existing literature on (a) the relationship between digital technologies and changing economic geographies, and (b) previous empirical approaches to measure platform work. The overview of the literature presented here is complemented by a more extensive literature review in the Supplementary Information (S3 Appendix: S3 Section, in particular S1 to S5 Tables in S1 File). In the appendix, we also provide more context on definitions and the conceptualisation of the term 'remote platform work' (S2 Appendix in S1 File). Here, we focus on a brief discussion of common findings, data, methods, and limitations of previous work from which we derive the contribution of our study.

The Economic Geography of Outsourcing, Offshoring, and Digital Technologies

The rise and role of ICT technologies for (global) economic geographies is a subject of extensive scholarly work and debate. Through outsourcing, offshoring and task automation, ICT technologies profoundly affect the location of economic activities and the global dynamics of interaction between economic centres. The internet has reduced frictions of economic interactions: low communication, transportation, and search costs, for example, enable more densely connected global value chains [50, 51].

Digital technologies facilitate the trade of commodities and in particular the trade of services at a distance. The offshoring of services allows for more complex global production networks [52] which further increase the international division of labour [53]. Several studies have attempted to capture the share and nature of jobs that could be offshored thanks to the increasing technological capabilities [34, 37, 38, 54]. While digital labour platforms are a relatively recent phenomenon, the offshoring of service activities has a long history. For example, in 1983, American Airlines established its first international back office in Barbados [55]. Paperwork was flown from the U. S., handled and digitised in Barbados and sent back electronically via satellite to save 50 per cent labour costs. In 2017, Feakins described how offshoring of services enabled complex global economic relationships, such as the Off-offshoring of service tasks from Russia to Ukraine [56].

ICTs are also reshaping the distribution of economic activities within countries [50]. For instance, Malecki (2002) observes that the geography of Internet backbone networks reinforces urban-rural differences [57]. Despite considerable controversy, the evidence overall suggests that the productivity impacts of ICTs are positive, but unevenly distributed [58, 59]. As a result, urban centres are likely to benefit disproportionately from digital technologies, as they tend tend to host specialised institutions and a more fine-grained division of labour than rural areas [22, 23, 60]. These are considered as enablers of complementary skills necessary to capitalise on the opportunities of digital technologies [61, 62]. The same holds for the integration of Global South countries into digital value chains. Overall, Hjort and Poulsen (2019) show that fast Internet access correlates positively with employment rates in African countries, due to the technology's impact on firm entry, productivity, and exports [15]. However, for example, Foster et al. (2018) find that East African firms involved in global value chains can only fully benefit from improved internet connectivity if they have access to complementary capacities and competitive advantages.

In summary, the internet has profoundly affected the geography of economic activities within and between countries across the globe over the past two to three decades. Overall, Information and Communication Technologies in general and the internet in particular have supported more fine-grained production chains and the offshoring of service activities. Online labour platforms continue this trend: they could represent one avenue for the further integration of remote workers into global digital value chains through outsourcing and offshoring. Studies from the platform economy literature have researched the geography of this novel form of online mediated work. In the following, we discuss the main findings and limitations of these studies with respect to the geography of remote platform work and the role of skills, and we derive our study's contribution from there.

The Economic Geography of Platform Work

Empirical studies on the geography of platform work have mainly focused on (a) assessments of the size of the platform labour market, (b) analyses of trade between countries, and (c) the role of signalling mechanisms and skills.

https://doi.org/10.1371/journal.pone.0274630

The first overarching finding is that platform labour markets are growing in importance and size [5, 13, 63] with estimated yearly growth rates of around 20% [6, 8].

Secondly, studies have investigated international transactions in the platform labour market: these are dominated by north-south interactions with employers from industrialised countries and workers from less developed countries [12, 13, 21, 63, 64]. Despite its inclusive global digital infrastructure, several barriers to trade and sources of worker discrimination persist such as geographical distance, language, time zone as well as cultural and ethical differences [13, 19–21, 64].

However, most of these studies use relatively small or old datasets, are limited to a small set of countries or investigations on the country-level. For example, Horten et al. (2017) [13] focus on the bilateral trade between countries in the online labour market with an economic gravity model. Their investigation aims on statistically explaining the large share of US-Indian trade on the online platform, and it stays largely on a descriptive level. Similarly, Agrawal et al. (2015) [63] provide descriptive statistics about the development of wages in occupational groups and selected countries over time, but do not explain the wage differences quantitatively. Lehdonvirta et al. (2019) [5] theorise online labour platforms through transaction cost economics and test their hypotheses empirically with a data set of 10,000 projects in two countries. Borchert et al. (2018) [65] and Braesemann et al. (2022) [23] analyse the relationship between online labour participation and local economic conditions on a sub-national level, but their investigations are limited to only one country.

These examples highlight a limitation of previous empirical contributions on the geography of platform work: many studies used relatively small cross-section data sets and they treated countries as the smallest unit of geographical analysis. This impeded a thorough identification of the economic determinants that shape the geography of online labour markets between countries and sub-national regions on a global level.

A third stream of research has focused on the role of skills in remote platform work. These studies find that while skills are considered as an important predictor of wages, workers seem to have limited opportunities to learn and grow on online platforms [12, 27, 28, 65]. Skill certificates can increase worker earnings [32]. However, there is contradicting evidence about the role of reputation systems in building trust, ranging from having an inclusive effect benefiting workers from developing countries disproportionately [21] to reputation leading to increasing inequality ("super star effect") [66]. Details of the platform design seems to play an important role.

In the studies on skills and signalling mechanisms, the operationalisation of skills represents a common challenge: many of the reviewed studies present simplistic operationalisations of skills or do not explicitly measure them at all. For example, Kässi and Lehdonvirta (2022) [32] focus on the wage effects of online workers obtaining skill certificates rather than on skill-based or occupational differences. Lehdonvirta et al. (2019) [5] compare only two occupations (writing vs. graphic design), similarly to Beerepoot and Lambregts (2015) [12] who operationalise differences in skill levels with only two occupations (web development = high skill vs. administrative support = low skill). Other studies focus more on reputation mechanisms than on skills or occupations as individual-level wage determinants [19, 66–68]. The network approach of Anderson (2017) [33] represents an exception. Anderson proposes a network-based method for measuring worker skills and finds that workers benefit from skill diversity.

In summary, while previous studies have investigated geographical and skill-related aspects of platform work, there is still a gap in the literature regarding the global distribution of online labour, and in particular its' relation to the unequal global distribution of skills, human capital, and place-bound institutions. We extend

the current literature on global geography of remote platform work by a quantitative analysis based on three relevant methodological advances: more and comprehensive data, geographic granularity, and matching of skills.

First, we collect one of the most extensive datasets on the subject, including almost 2 million projects spanning eight years from 2013—2020 and a fine-grained geographical granularity. Secondly, we take a global viewpoint including 139 countries while extending the analysis to the sub-national level. In particular, we map the geocoded online platform data to regional statistical data from the World Bank, the OECD and the Global Data Lab on the country- and sub-national level, which makes the global persistence of urban-rural differences visible for the first time. This step also allows for applying regression modelling to explain the geographical distribution in relation to local economic conditions and infrastructure. Thirdly, we propose a sound methodology to operationalise skills by matching remote jobs to a well-established occupational taxonomy. These three methodological advances allow us to pose the following research questions:

RQ 1 How is remote platform work globally distributed between and within countries? What is the relationship between urban and rural areas?
RQ 2 Are the unequal global distributions of remote platform work activity and wage differentials in Global North and Global South countries determined by local institutions, infrastructure and economic conditions?
RQ 3 In how far are outcomes of different job types in the remote platform labour market (wage levels, value of experience) related to differences in competitive intensity and skill requirements across occupations?

Data and Methods

All the methods of the data collection and analysis are described in great detail in the Supporting Information (S4 and S5 Appendices in S1 File). Here, we summarise the most critical steps.

Data Collection

In this study, we combine three data sources: (a) transaction records from a globally leading online labour platform, (b) regional covariates covering the demography, economy and infrastructure in OECD+ and Global South countries, and (c) occupation statistics from the U. S. Bureau of Labour Statistics (S4 Appendix in S1 File). The retrieval and assembly of online labour records proceeds in two steps: After having gathered information via the API about the projects of which we had IDs, we extracted the platform worker IDs from these projects. In total, we collected a data set of 1.8 million remote jobs from one global online labour platform covering eight years from 2013 to 2020.

Geocoding

In a second step, we provided these IDs to the API to obtain the remaining information related to each transaction of these platform workers. This includes the hourly wage, the total price charged for the project, and the workers' country-city location (S4 Appendix: S4.1 Section in S1 File). Afterwards, we use a Geocoding API and provide it with a list of all unique country-city locations from both the employer and worker side of the platform transactions; a total of 66,085 locations (S4 Appendix: S4.2 Section in S1 File). Then, the geocoded online labour data is matched with national and sub-national statistics on demography, economy, and infrastructure. Here, three data sources are considered: World Bank for country-level statistics, OECD regional statistics for sub-national data in high and middle income countries from the Global North,

https://doi.org/10.1371/journal.pone.0274630

and Global Data Lab for sub-national level data for low and middle income countries from the Global South (S4 Appendix: S4.3 Section in S1 File).

Occupation Mapping

Besides the geographical analysis of online labour data, we also investigate the job types of the online projects. For this purpose, we match the online job categories with official occupational statistics used by the U. S. Bureau of Labour Statistics (BLS). The BLS provides detailed information about each occupation's educational requirements, skills, and abilities. This data is available via the Occupational Information Network O*NET (S4 Appendix: S4.4 Section in S1 File). To match online work descriptions with official occupational taxonomies, we use the SOCcer (Standardized Occupation Coding for Computer-assisted Epidemiological Research) tool provided by the U. S. National Institutes of Health for an automatised coding of a sample of 345,000 online projects. We provide the online job category as job title, and the required skills and description of the online project as job description to the tool. Based on the occupational mapping, we derived two measures 1) capturing the skill or educational requirements of different occupations from BLS data and 2) the relevance of experience in obtaining online projects (S4 Appendix: S4.5 Section in S1 File). We present the distribution of skill requirements and use hierarchical cluster, applying a Euclidean distance measure and complete linkage as clustering method, to group skills and occupations. Furthermore, we relate occupation-level variables to the importance of experience in obtaining online projects. This is what we call the 'experience gradient'. The idea is that experience and reputation are known to drive outcomes in the platform economy, as they signal trustworthiness of sellers.

Regression Analysis

Lastly, the regression analysis of the geographical distribution of online labour projects and wages relies on six regression models with multilevel effects, where we regress a broad set of regional characteristics on regional- and country-level wages and project count whilst ensuring accurate feature selection and out-of-sample prediction accuracy (S5 Appendix in S1 File), including various robustness checks to ensure that our findings are consistent across time and space (S6 Appendix in S1 File).

Results

Polarisation Across Space

Considering Research Question 1 on the global distribution of remote platform work, the data shows that the online labour market connects global demand for and supply of remote knowledge work (Fig 2.1). However, while it is theoretically open to users from all over the world, Fig 2.1A shows that demand and supply highly cluster in a limited number of places. Most demand comes from urban areas in North America, West Europe, and Australia; most remote platform workers are located in East Europe, South Asia, and the Philippines. Dense flows of capital and labour connect these regions, while many Global South countries only marginally participate in the remote labour market. The overall polarisation remains largely persistent over time (S6 Appendix: S6.5 Section in S1 File).

https://doi.org/10.1371/journal.pone.0274630

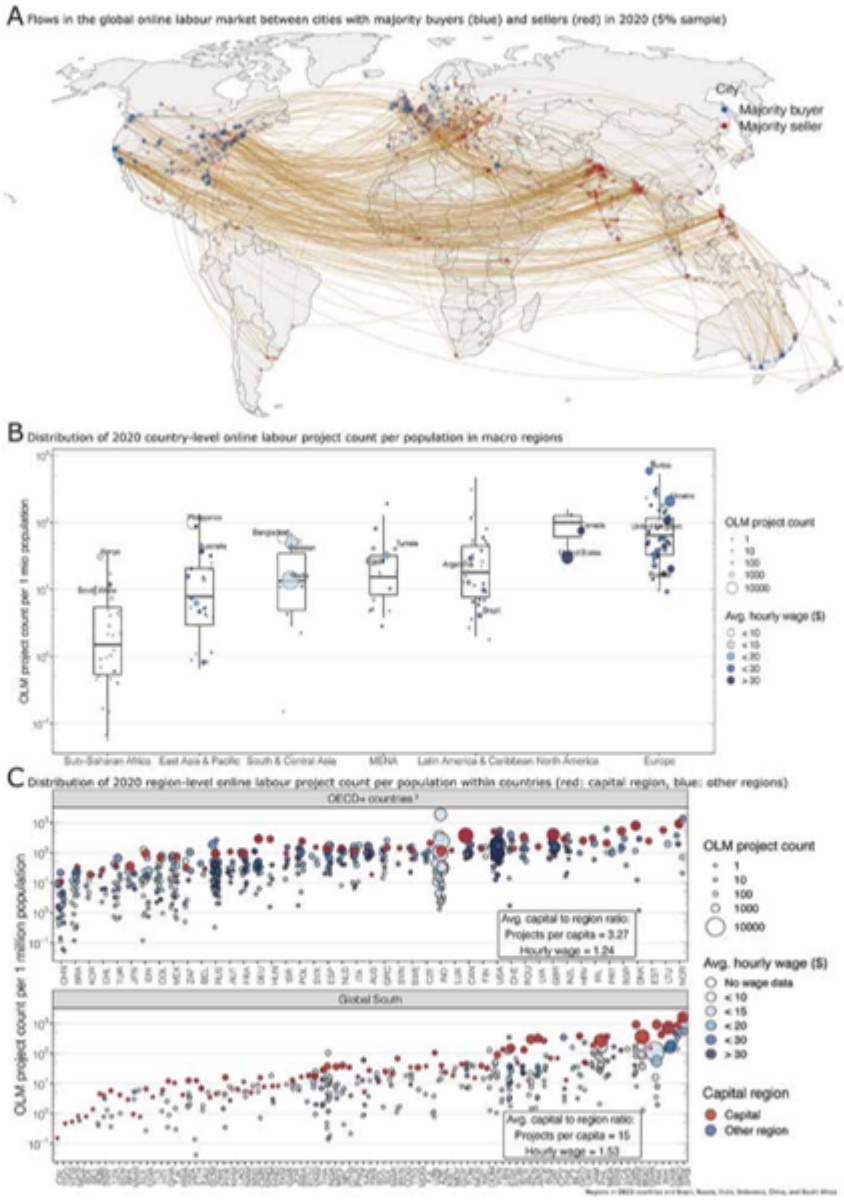


Fig 2.1 (**A**) Connections between majority buyer (blue) and seller cities (red) in the 2020 platform labour market (5% sample): hotspots of demand are North America, West Europe, and Australia; platform workers in Eastern Europe, South Asia and the Philippines conduct most remote jobs. (**B**) Distribution of 2020 online labour (OLM) project count per capita (y-axis) in countries (dots), grouped by global macro regions (x-axis): globally, platform activity varies by several orders of magnitude; Europe and North America show the highest levels of participation, and the highest average wages (dot colour); most countries in the Global South participate only marginally in the remote labour market with low wages and less than 10 projects per one million population, with the exception of the Philippines, Bangladesh, Pakistan, and India. (**C**) Online labour distribution within countries in OECD+ and Global South countries: participation varies vastly within countries with most capital regions (red) hosting the largest platform worker communities per country; the imbalance is particularly pronounced in the Global South where the capital region hosts, on average, 15-times as many projects per capita than other regions in the country.

https://doi.org/10.1371/journal.pone.0274630.g001

The global differences in participation become more pronounced when considering the number of projects per capita (Fig 2.1B). Online labour project count per population varies by several orders of magnitude within and between macro-regions (position of the dots on the y-axis): while almost all countries in Europe and North America hosted at least ten projects per one million population in 2020, only half of the countries in South & Central Asia, one-third in East Asia & Pacific, and 15% in Sub-Saharan Africa did so. In absolute numbers (size of the dots), more than 50% of all online labour projects were conducted by platform workers from just five countries (India, Pakistan, Philippines, United States, Bangladesh). Hourly wages (colour of the dots) also vary substantially: while platform workers in the United States, United Kingdom, and Russia charged more than 30 USD per hour on average, remote workers in Bangladesh and the Philippines earned just a fifth, or 6 USD an hour. There are also exceptions: for example, Kenya and South Africa host relatively active platform worker communities. Kenya's participation per population is comparable to that of the United States. Averages wagers of South Africa's platform workers are comparable to those in Europe.

Inequalities between countries online labour activity have been reported in the past [13, 63]. However, here we also reveal the sub-national concentration of remote platform work globally (For more details, see S6 Appendix: S6.3 Section in S1 File). Participation rates vary by two to four orders of magnitude in many countries (Fig 2.1C), and the distribution within and between countries is highly concentrated, both in OECD + BRIICS (OECD countries and Brazil, Russia, India, Indonesia, China, and South Africa, abbreviated as OECD+). and Global South countries (S6 Appendix: S17 Fig in S1 File). In many countries, the capital region attracts most platform jobs in absolute and per capita terms (Fig 2.1C), particularly in the Global South. In the OECD+, the capital regions attract more than three times as many platform jobs per capita than other regions in the same county on average. In the Global South, capital regions obtain more than 15 times as many projects per capita as other regions in the same country. On a global scale, platform work is a metropolitan phenomenon.

Additionally, hourly wages vary substantially between metropolitan and other regions. On average, platform workers in OECD+ capital regions earned 24% more per hour than their counterparts in other regions. The wage spread was almost twice as high in Global South countries: platform workers in capital regions earned 53% more than those in other regions.

Summarising the findings related to Research Question 1, the data suggest two dimensions of geographical polarisation in the global remote labour market. First, we find pronounced inequalities between countries worldwide along a North-South dimension. Most online labour demand comes from high-income countries. Supply comes mainly from platform workers in traditional outsourcing destinations in South Asia, the Philippines, and by platform workers from Europe and North America. Secondly, we identify persistent inequalities within countries. This points towards urban-rural differences as the second main polarisation dimension in the remote labour market. Platform workers in large cities and places with enabling institutions seem to be able to secure more and better paid jobs than their counterparts in rural regions.

Turning to Research Question 2, we find that regional factors explain the spatial polarisation described in Fig 2.1. Fig 2.2 shows six regression models relating the number of projects and hourly wages per country (models 1 and 4), per OECD+ region (models 2 and 5), and per Global South region (models 3 and 6) to country-level and sub-national covariates. Data sources are World Bank data (countries), the OECD regional database (sub-national regions OECD+), and the Global Data Lab [69] (sub-national regions Global South). The data covers the years 2013—2020 (for details on the data set and pre-processing, see S4 Appendix: S4.1–S4. sections and S5 Appendix: S5.1 Section in S1 File). In summary, the regression models contain data of 139 countries an 597 sub-national regions over eight years.

https://doi.org/10.1371/journal.pone.0274630

A Regression models relating online labour project count (1–3) and avg. hourly wage (4–6) to regional covariates 2013 to 2020

Dependent variable:	Yearly online labour project count[a]			Online labour avg. wage per hour[a,b]		
Level:	Countries	Sub-national regions		Countries	Sub-national regions	
Geography:	Global	OECD+[c]	Global South[d]	Global	OECD+[c]	Global South[d]
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Population Population, total (log scale)	0.96*** (0.02)	0.85*** (0.09)	0.54*** (0.04)	0.04*** (0.01)	0.05 (0.04)	0.06*** (0.02)
Education Model (1), (4): share of pop. with secondary education Model (2), (5): share of pop. with tertiary education Model (3), (6): avg. years of education	0.06*** (0.003)	0.004 (0.01)	0.07*** (0.03)	0.01*** (0.001)	0.01*** (0.002)	0.06*** (0.02)
Income per capita Model (1), (4): GDP per capita (in 1,000 $) Model (2), (5): GDP per capita (2015 PPP $, log scale) Model (3), (6): Gross National Income p.c. (2011 PPP $)	−0.01*** (0.003)	−0.33*** (0.10)	0.06*** (0.02)	0.002*** (0.001)	−0.12*** (0.05)	0.03*** (0.01)
Internet connectivity Model (1), (4): fixed broadband subscriptions per 100 people Model (2), (5): share of HHs with internet broadband access Model (3), (6): share of HHs with internet access	0.04*** (0.01)	0.03*** (0.004)	0.02*** (0.004)	0.01*** (0.002)	0.005*** (0.001)	0.003** (0.001)
IT specialisation of the economy Model (1), (4): ICT share of all service exports (log scale) Model (2), (5): Gross value added in ICT (2015 PPP $, log)	0.22*** (0.04)	0.56*** (0.05)		−0.02* (0.01)	0.09*** (0.02)	
English language Indicator: English is official language	0.68*** (0.11)			−0.07** (0.03)		
Price level PPP conversion factor (per 1,000 int. $)	−0.32*** (0.08)			−0.04 (0.02)		
Capital region Indicator: region holds country capital		0.13 (0.10)	1.75*** (0.10)		−0.08** (0.03)	−0.12** (0.05)
Constant					3.21*** (0.31)	2.49*** (0.12)
n (regional units)	139	292	305	112	253	56
Observations	1,136	2,384	2,122	763	1,536	255
Fixed / Random Effects	Yearly FE	Country-year Fixed Effects		Yearly FE	Country-year Random Effects	
R^2	0.72	0.70	0.43	0.44	0.79	0.69
Adjusted R^2	0.71	0.67	0.36	0.43	0.79	0.68

Note: *p<0.1; **p<0.05; ***p<0.01

[a]Inverse hyperbolic sine (ihs) transformed: $y = \log(x + \sqrt{x^2 + 1})$.

[b]To ensure convergence of the mean wage, countries or regions with a project count of less than 25 have been excluded from the regression models (4)–(6).

[c]Regions in OECD countries and Brazil, Russia, India, Indonesia, China, and South Africa.

[d]Regions in Global South countries available in the *Global Data Lab* database.

B Spread of online labour project count (left panel) and average hourly wage (right panel) per country in 2020, compared to the spread of the residuals of the optimised models (1) and (4)

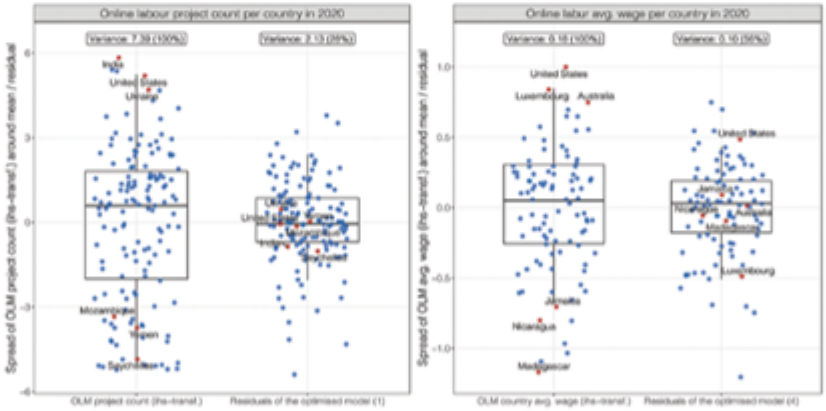


Fig 2.2 (**A**) Regressions between online labour project count (models 1—3), avg. hourly wage (models 4—6) and regional covariates (2013—2020 data, in total 1.76 million projects): population, education, internet connectivity, and the IT specialisation of the economy are positively associated with project count and hourly wages; globally, countries with English language and low price levels are more active in the remote labour market. (**B**) Spread of online labour (OLM) project count (left panel) and avg. hourly wage (right panel) per country vs. residuals of the regression models (1) and (4): the parsimonious models explain large shares of the global variation; for example, the countries at both ends of the project and wage spectrum (highlighted in red) show substantially reduced residuals after controlling for regional covariates. Overall, the regression models explain between 42% and 79% of the variation between countries or regions.

https://doi.org/10.1371/journal.pone.0274630.g002

The dependent variable in the models (1) to (3) is inverse hyperbolic sine (ihs) transformed $\left(y = log(x + \sqrt{x^2 = 1})\right)$, see [70]) to reduce the skewness of the distribution, which ranges across several orders of magnitude. Hourly wages also vary substantially. Therefore, we also applied the ihs-transformation to the dependent variable in the models (4) to (6). As our models deal with different hierarchical levels (regions nested in countries or years), multi-level effects need to be considered. We test and apply random and fixed effects to account for the variability of outcomes within and across countries or years. (see S5 Appendix: S5:3 Section in S1 File).

To model the relation between the platform data and regional characteristics, we included regional statistics commonly used in studies on the platform economy (see S3 Appendix, S4 Appendix: S4.3 Section, and S5 Appendix: S5:2 Section in S1 File).

The regression models (1) to (3) tell a coherent and robust story about the geography of the platform labour market. The larger a region's population, the more projects it attracts. Similarly, higher levels of education are associated with higher project counts. The income level is negatively associated with project count on the global scale and in the OECD+, while it is positively associated in Global South regions. This indicates that it is middle income countries or regions, not the poorest places on earth, that attract most remote jobs. Internet connectivity is positively associated with project count: the better the internet infrastructure, the more remote platform work. The same holds the 'IT specialisation of the economy'. English language countries attract more projects, as well as those that have a lower price level. In Global South regions, the capital region indicator is strongly positively associated with activity.

The models (4) to (6) give insights into the factors that drive hourly wages. Overall, the coefficients show a similar direction as in the models (1) to (3), with a few important exceptions. For example, the English language coefficient is negative in model (4) because many high-wage countries in Europe do not have English as an official language. The 'IT specialisation of the economy' coefficient is negative in model (4). Income is positively associated on the country level and the price level is not significant (potentially because platform workers in countries with higher price levels tend to charge more). Education and internet connectivity are positively correlated with hourly wages.

The regression models explain a large share of the total variation (see R^2 values in Fig 2.2A and 2.2B). Almost three-quarters of the total variance between countries (left panel of Fig 2.2B) can be explained by the model: the residuals of very large players in the online labour market, such as India, the United States, or Ukraine (red dots) are almost zero. Similarly, model 4 accounts for 44% of global wage differentials (right panel of Fig 2.2B).

In summary, we conclude that the global distribution of remote work seems to be constrained by place-bound economic, infrastructure, and educational factors (RQ 2). The most profitable projects are conducted in places with high human capital levels, specialised know-how, and a robust local economy. Price differentials explain only a minor share of the global geography (S6 Appendix: S6.2 Section in S1 File).

Polarisation Across Skills

To explain the wage differentials between job types in the platform labour market (Research Question 3), we look into the skills related to the professions. Following the task-based approach [34, 37, 71], we consider jobs as the manifestation of different tasks and skills. Fig 2.3A displays the skill composition of each of the 46 occupations the platform job types can be grouped in (see S4 Appendix: S4.5 Section in S1 File for details on the matching between online job types and official occupational categories). Skills (rows) and occupations

(columns) are sorted by a hierarchical clustering algorithm (see S4 Appendix: S4.5 Section in S1 File) to group occupations with similar skill requirements, resulting in nine skill- and six occupation cluster.

The upper half of the heatmap reveals significant differences between the occupations. Job types in clusters 3 and 4 score heavily on computer-related skills; those in clusters 2 and 6 score most intensively on language and (written) communication skills. Occupations in clusters 1 and 5 score most heavily on oral communication or clerical tasks.

The varying skill requirements translate into different hourly wages (lower part of Fig 2.3A). Jobs in clusters 3 and 4 pay an average wage of $ 16—17, jobs in cluster 2 and 6 an average wage of $ 15—16, and jobs in cluster 1 and 5 a mere average of $ 6—9 per hour. In other words, skill sets determine wages. This is confirmed by two regression models in Fig 2.3B). Model (1) relates the average hourly wages per occupation to three variables: the average number of applicants per project (competitive intensity), the total number of projects per occupation (size of each occupational sub-market), and the educational attainment score (reflecting differences in required education levels, see S4 Appendix: S4.5 Section in S1 File). The model identifies a strong negative association between wages and competitive intensity. The educational attainment score is positively associated with wages, particularly for 'Non-Tech' occupations (see panel (i) and (ii) in Fig 2.3C).

For platform workers, expected wages are essential but not the only relevant outcome. Due to uncertainty in the remote labour market, workers need to send quality signals to demonstrate experience and trustworthiness to potential employers, such as ratings or reviews about past projects [67, 72]. Platform workers that can secure some initial projects to distinguish themselves from other, less experienced, competitors might be able to obtain more profitable projects in the future. To operationalise the varying importance of experience signals in different occupations, we developed a statistical measure of the relevance of past experience in obtaining additional projects: the *experience gradient*. We calculate the experience gradient per occupation as the slope parameter estimate $\hat{\beta}$ of a regression between the number of projects a platform worker in occupation i obtained in year t by the number of projects the same worker had conducted in previous years (see S4 Appendix: S4.5 Section in S1 File). The experience gradient is related to occupation-level variables in model (2) of Fig 2.3B. The model shows that competitive intensity, measured by the market size, and average wages are positively associated to the relevance of experience. This applies more strongly to Non-Tech jobs (panels (iii) and (iv) in Fig 2.3C): in more competitive occupations with less skill-based signalling options, experience is more relevant than in others.

Interpretation of the Results

Here, we provide an interpretation of the quantitative findings regarding the spatial and skill-based polarisation or remote platform work. With respect to the geographical distribution (Research Questions 1 and 2), we argue that remote work does not simply flow to places with lower price levels, but that it is pulled towards locations with competitive advantages and a specialisation in information and communication technologies. This could be related to centripetal forces that characterise the spatial organisation of the institutions enabling remote knowledge work. These forces seem to be bigger than global price differentials, which could push remote work to places with low living costs. The places with the lowest price levels globally tend to not have a sufficient level of internet infrastructure, economic specialisation and know-how to enable workers to participate in the global market for remote knowledge work. The global polarisation in the remote labour market is then a digital mirror of the global polarisation of skills and economic opportunities across the globe.

https://doi.org/10.1371/journal.pone.0274630

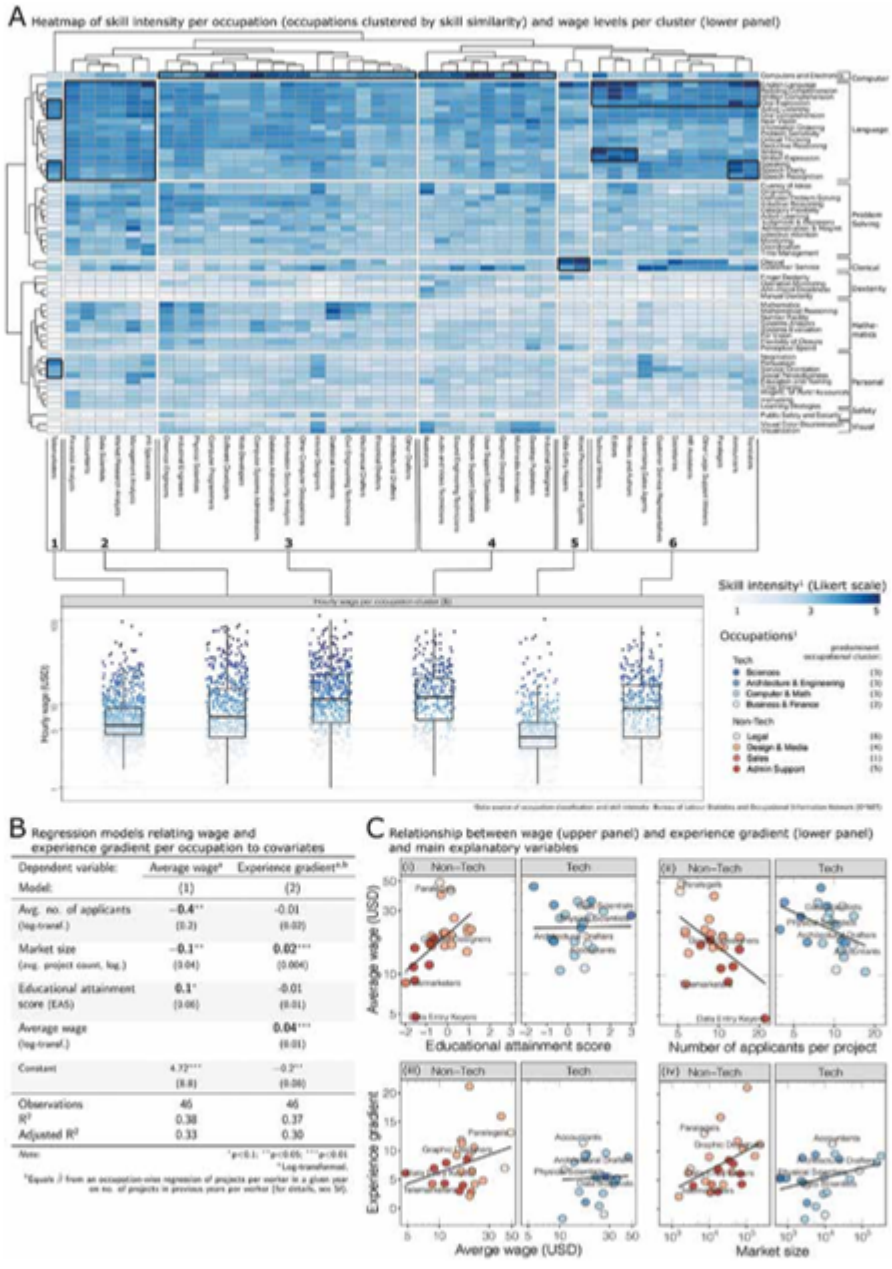


Fig 2.3 (**A**) Heatmap of skill intensity per occupation (upper panel) and wages per occupation cluster (lower panel): occupations (columns) with similar skill intensities (rows) cluster together; the highest paying occupations require computer-related know-how or English language comprehension and writing; lower-paying occupations focus on clerical skills, personal and oral communication. (**B**) Regression models between hourly wage (model 1), experience gradient (model 2) and occupation-level covariates: occupations with less competition (fewer applicants per job and lower avg. project count) and higher educational requirements pay higher average wages; previous experience counts more in high-paying occupations with fiercer competition (high project count). (**C**) Relations between wage (upper panel), experience gradient (lower panel) and occupation-level covariates: platform workers with only a fewer previous projects find it hard to be hired in non-tech occupations (panel iv).

https://doi.org/10.1371/journal.pone.0274630.g003

Considering the results with regards to the distribution of skills and wages in the online labour market, we conclude the following. First, in contrast to conventional labour markets, which are shaped by geographical and regulatory constraints leading to substantial wage differentials for similar types of jobs even in close geographical proximity—think of wage differentials at the US-Mexican border region, Hong Kong vs Mainland China, Switzerland vs adjacent European countries or South Europe vs North Africa—the remote platform labour market is truly global. It is, however, not just one market, but many: one market per occupation. The variation in hourly wages between occupations is larger than the overall variation between countries (S6 Appendix: S6.4 Section in S1 File).

Secondly, the regression models show that jobs with more applicants and less skill-based signalling pay lower wages (Research Question 3). We interpret this finding as follows: These occupations have low entry barriers and face more competition. Without skills as a quality signal, the number of prior projects or feedback scores becomes an entry barrier. Employers use trust cues to decide whom to hire from the large crowd of applicants. For platform workers, this could spark a race to the bottom: without reputation, they find it hard to get their first job, so they need to undercut wages, leading to a vicious circle of more competition and lower wages.

In contrast, people with in-demand skills can secure profitable projects and high wages. The skill-based polarisation does not work along a one-dimensional skill axis: for example, the highest paying occupation is 'Paralegals and Legal Assistants'. This occupation does not require an exceptionally high level of formal education, but knowledge of the legal system in the country of the employer (in many cases the United States). Similarly, 'Announcers' (i. e. audio online adverts etc.) receive high wages. This occupation comes with another hard-to-copy skill: an U. S. accent. In general, jobs with hard-to-acquire, technical skills show less competition and higher wages.

Overall, the findings lead to the conclusion that demand for and supply of skills drive outcomes in the remote labour market. The three axes of polarisation—global divides between countries, urban-rural imbalances within countries, and inequalities between occupations—reflect the scarcity and abundance of skills and the access to them. In the global platform labour market, the laws of supply and demand work unrestrained: individuals with in-demand skills are able to secure profitable jobs; others obtain low wages, face fierce competition, and reputation as a crucial entry barrier. The outcomes of individual platform workers are constrained by system-level mechanisms shaped by agglomerative forces, which are largely out of their hands. The jobs remote workers can perform online is then determined by their access to education, training, and specialised IT know-how. This access is linked to place-bound institutions of the local economy. If they are unlucky not to be located near specialised industries or agglomerations, they are more likely to offer work in occupations characterised by easy-to-copy skills and fierce competition. In contrast, remote workers from metropolitan areas already have access to ample urban opportunities for knowledge exchange and local work opportunities, due to a bundling of complex economic activities [22] and a more fine-grained division of labour in urban environments [60]. They will enjoy the increased global demand for IT and business services and attractive wages.

Discussion

The Covid-19 pandemic has led to the rise of remote work. Digital technologies enable the practical organisation of work at a distance. The potential cost savings, more flexibility, and improved access to talent suggest that remote work is likely to play an essential role in the future of work. However, it is unclear how far remote work will influence the global division of labour. In particular, does remote work bring jobs to rural areas and disadvantaged regions or will it reinforce global spatial imbalances? To investigate this question quantitatively, we draw on data from a fully remote labour market: an online labour platform. In the empirical part of this study, we analyse remote work mediated by an online labour platform. Here, in the

https://doi.org/10.1371/journal.pone.0274630

discussion section, we also consider what implications the quantitative findings may have for remote work more generally, including regular employment performed remotely over the internet.

On the platform, workers from all over the world can find and conduct jobs covering the whole spectrum of knowledge work. As the whole work process—from the job advert over the interview, onboarding, communication, to payment and dispute resolution—is conducted online, the platform labour market could provide an outlook into the future of work. This future might be shaped by more remote contracts [73, 74] and the platformisation of jobs [75–77].

The data suggest that agglomerative forces shape the global geography of remote work, leading to the following interpretation. Jobs are pulled to places with enabling institutions of remote work, which are unequally distributed across the globe. Complex economic activities and specialised vocational training concentrate in large cities [22, 42]. In having access to these opportunities and skills, the remote workforce in urban areas is able to obtain the most profitable remote jobs, while their counterparts outside of economic centres find it more difficult to offer in-demand skills on the global platform labour market. The unequal spatial distribution of skills, institutions, and opportunities determine the global geography of remote work. The dynamics of the platform economy amplify this process, as there are little geographical or regulatory boundaries on the online platform that would slow down the global competition.

Across countries, we observe a spatial division of work that resembles the offshoring rationale of business processes, which started in the 1980s and 1990s [4]. Increasingly modularised and standardised tasks within the ever-growing digital economy have enabled a fine-grained global division of knowledge work connecting North America, West Europe, and Australia with South Asia, the Philippines, and East Europe. This observation is in accordance with earlier work that investigated the increasingly globalised market for knowledge work and Digital Taylorism—the modularisation and standardisation of cognitive tasks together with a fine-grained division of these tasks across different job types in global digital value chains—such as 'The Global Auction' by Brown et al. (2010). However, the data shows that most countries in the Global South are only marginally connected to the global web of remote work in the platform labour market.

Within countries, we find that remote work flows to urban centres. These are the places where highly skilled labour is concentrated. The economic tale of the 'booming metropolis' and the 'broken provincial city' [40] plays out fully in the platform economy. This imbalance resembles existing opportunity gaps between urban and rural areas, which have pulled skilled workers to cities already before the rise of the platform economy. In fact, while online labour platforms are thought to offer an alternative to employment-based migration, constituting a form of 'virtual migration' [78], studies suggest that online labour platforms are frequently used by migrants [79]. Similarly, online labour platforms are more intensively used by younger parts of the working age population. Pajarinen et al. (2018) find that many online freelancers in Finland are less than 30 years old [79]. It is the young, talented part of the workforce that most likely migrates to urban areas because of better income opportunities [80]. Such opportunity-based migration to metropolitan areas and resulting differences in the age structure of urban and rural areas is probably one of the drivers of the geographical disparities in the platform data. Overall, we observe that remote platform workers in metropolitan areas are more likely to attract specialised online jobs and high wages.

The findings highlight the pivotal role of skills in driving remote work towards metropolitan areas. Individual occupations form sub-markets of the global platform labour market. Platform workers are constrained to work in those jobs that reflect their skills and experiences [81]. Competitive pressure differs between job types as workers cannot freely move between occupations. This leads to a scarcity premium in some occupations, while others suffer from low wages and excess supply. In that situation, the uncertainties of the platform economy spark a race to the bottom: feedback and prior work experience are highly relevant to obtain remote jobs [46, 82]; newcomers will find it hard to get their first job and might be forced to undercut wages. This

https://doi.org/10.1371/journal.pone.0274630

fuels competition and spirals wages downwards. Besides the intense competition and limited opportunities for signalling work quality, which we describe in this study quantitatively, the literature has discussed the role of algorithmic control [27, 83] and the organisation mechanisms of work in the platform economy [30, 84] as drivers of (adverse) outcomes for remote platform workers.

Our analysis implies that agglomerative forces drive the polarisation of the remote labour market. Market access alone will not lead to a less unequal division of labour. Remote workers in disadvantaged areas need more than a computer and broadband internet alone to thrive. They need to have marketable skills to make remote work a tool for rural and economic development.

Policy Implications

Initiatives such as the Rockefeller Foundation's *Digital Jobs Africa* or Kenya's *Ajira Digital* work programme aim to bring millions of remote jobs to Africa, but they could make matters worse for remote workers: if they increase the supply in certain types of occupations, they could fuel the competitive spiral of excess labour supply and pressure on wages. To increase chances of remote workers in disadvantaged regions, retraining programmes need to focus on in-demand skills and account for the quickly changing dynamics in the global market for talent. It is unlikely that remote worker communities can thrive if there are limited local opportunities. Therefore, online work programmes in rural areas—both in Global North and Global South countries—should be embedded in larger economic and labour market development schemes, which provide reliable internet access, local employment alternatives, and skill-building opportunities. This applies also to remote labour demand. Remote platform work can be a chance for rural employers to get access to global talent. Programmes that foster the integration of remote work into their business processes might help companies to become more resilient and to keep them in their local surrounding.

Online platform providers could increase the visibility of objective quality metrics, such as educational degrees, to limit the adverse effects of reputational feedback loops in the remote platform labour market. Platform apprenticeships for new remote workers—the random assignment of first jobs to people without experience on the platform—could help to build up initial credibility [67] and lower entry barriers. Moreover, governmental organisations that aim to improve the working conditions of remote platform workers, such as the European Commission, could support the positive development of platform work in advertising short-term remote jobs directly on online platforms while promoting living wages. They could also help in developing and accrediting objective quality metrics. Another way to strengthen local communities of remote platform workers could be the support of coworking spaces and other forms of physical meeting points for platform workers. Such spaces could focus on providing workplaces with all the equipment and services necessary for performing remote jobs effectively. In bringing remote workers together and offering complementary services, these spaces could foster knowledge exchange and skill-building, and they could help remote platform workers to gain a sense of community.

Implications for the Future of Remote Work

Why is it that remote work is unlikely to change the economic imbalance between urban and rural areas majorly? Remote work allows people to move freely from urban to rural areas only in the short term. Some urban specialists might enjoy the new remote work opportunities and relocate to suburban or rural environments, substituting day-to-day commutes with digital interactions. They can stay connected to their peers in urban centres via video calls, but they will find it difficult to establish new links and access informal knowledge exchange through local networks.

In contrast, the city, as a hub of interaction supports specialised local jobs and occupational diversity [60, 85, 86]. An increasing share of complex economic activities [22], and non-linear scaling laws make the provision

of specialised occupations more sustainable in large agglomerations [87–89]. This will likely continue to pull business opportunities and people towards cities. Digital technologies and organisational adjustments during the Covid pandemic have enabled seamless communication and collaboration over distance, which theoretically allows for a wide-spread web of economic and labour market interactions between urban and rural environments. Still, the network forces that pull innovation and business opportunities towards large agglomerations are likely to also shape the global geography of remote work. In the remote labour market, the place of work might not be limited to the same city anymore. But instead of dispersing more equally across space, remote work probably tends to cluster in metropolitan areas in different parts of the world, which share similar institutions and urban lifestyle that enable knowledge work to flourish locally.

Limitations

Our study comes with some methodological limitations. The data collection (described in S4 Appendix in S1 File) dependents on access to the online platform's API. We cannot make claims about the size of our sample in relation to the overall size of the remote platform labour market, but we are confident that sampling issues did not bias the analysis. This is because our findings are robust over all the years in the sample and they align with previous investigations on the geography of platform work. Furthermore, Our study analyses data from only one platform, but the platform investigated here is one of the global market leaders. In the data analysis, we had to make simplifying assumptions: in mapping the platform job categories to the official occupation taxonomy, we had to disregard the multifaceted skill-dimensionality of jobs within each occupation. Moreover, the algorithmic geocoding and occupation mapping come with some uncertainties. However, we very carefully investigated each step of the data preparation for potential errors (outlined in the SI) and we performed several robustness checks (for example, S6 Appendix: S6.1 Section in S1 File) to validate parameter choices.

Conclusion

Remote work mirrors the spatial inequalities of labour markets at large. The most profitable jobs are pulled towards the booming tech-savvy metropolis, while rural areas fall behind. In contrast to on-site labour markets, the polarisation mechanisms are amplified in the platform economy, as the forces of supply and demand are fully unleashed in the absence of regulatory barriers.

The unequal spatial distribution of institutions enabling skill-building and business opportunities determine the geography of the remote labour market. While internet connectivity and price differentials channel remote work around the globe, only remote workers with access to specialised skills attract valuable projects. Platform reputation mechanisms further accelerate the global race to the bottom for those who do not possess in-demand skills. Still, remote work can become an instrument of economic empowerment and growth. For this to happen, remote work needs to be embedded in broader economic and labour market development schemes, supporting disadvantaged regions to invest in local skill development and infrastructure.

Only in regions that flourish locally, remote workers can succeed globally.

Supporting Information

S1 File.
(PDF)

S1 Text.
(PDF)

https://doi.org/10.1371/journal.pone.0274630

Author Contributions
Conceptualization: Fabian Braesemann, Mark Graham.
Data curation: Fabian Braesemann, Otto Kässi.
Formal analysis: Fabian Braesemann, Fabian Stephany, Ole Teutloff.
Funding acquisition: Fabian Braesemann, Mark Graham, Vili Lehdonvirta.
Investigation: Fabian Braesemann, Fabian Stephany, Ole Teutloff, Vili Lehdonvirta.
Methodology: Fabian Braesemann, Fabian Stephany, Otto Kässi, Vili Lehdonvirta.
Project administration: Fabian Braesemann.
Resources: Fabian Braesemann, Otto Kässi, Mark Graham.
Software: Fabian Braesemann, Fabian Stephany.
Validation: Fabian Braesemann, Fabian Stephany, Ole Teutloff, Otto Kässi, Mark Graham, Vili Lehdonvirta.
Visualization: Fabian Braesemann, Fabian Stephany.
Writing – original draft: Fabian Braesemann, Fabian Stephany, Ole Teutloff, Otto Kässi, Mark Graham.
Writing – review & editing: Fabian Braesemann, Fabian Stephany, Ole Teutloff.

References

1. Forman Chris and van Zeebroeck Nicolas. From wires to partners: How the internet has fostered r&d collaborations within firms. *Management science*, 58(8):1549–1568, 2012. https://doi.org/10.1287/mnsc.1110.1505
2. Carnevale Joel B and Hatak Isabella. Employee adjustment and well-being in the era of covid-19: Implications for human resource management. *Journal of Business Research*, 116:183–187, 2020. https:// doi.org/10.1016/j.jbusres.2020.05.037
3. Brodie Boland, Aaron De Smet, Rob Palter, and Aditya Sanghvi. Reimagining the office and work life after covid-19. White paper, McKinsey, 2020.
4. Manning Stephan, Larsen Marcus M., and Chacko G. Kannothra. Global Sourcing of Business Processes. In *The New Oxford Handbook of Economic Geography*, pages 407–426. Oxford University Press, 2018.
5. Vili Lehdonvirta, Otto Kässi, Isis Hjorth, Helena Barnard, and Mark Graham. The global platform economy: A new offshoring institution enabling emerging-economy microproviders. *Journal of Management*, 45(2):567–599, 2019.
6. Otto Kässi, Vili Lehdonvirta, and Fabian Stephany. How many online workers are there in the world? a data-driven assessment [version 2; peer review: 3 approved. 2021.
7. Richard Heeks. Decent Work and the Digital Gig Economy. Development Informatics Working Paper 71, SSRN, 2017.
8. Kässi Otto and Lehdonvirta Vili. Online labour index: Measuring the online gig economy for policy and research. *Technological forecasting and social change*, 137:241–248, 2018. https://doi.org/10.1016/j.techfore.2018.07.056
9. Stephany Fabian, Dunn Michael, Sawyer Steven, and Lehdonvirta Vili. Distancing bonus or downscaling loss? the changing livelihood of us online workers in times of covid-19. *Tijdschrift voor economische en sociale geografie*, 111(3):561–573, 2020. https://doi.org/10.1111/tesg.12455
10. Thompson Beverly Yuen. The digital nomad lifestyle:(remote) work/leisure balance, privilege, and constructed 661 community. *International Journal of the Sociology of Leisure*, 2(1):27–42, 2019.

11. James Manyika, Susan Lund, Kelsey Robinson, John Valentino, and Richard Dobbs. A labor market that works: Connecting talent with opportunity in the digital age. Working paper, McKinsey Global Institute, 2015.

12. Niels Beerepoot and Bart Lambregts. Competition in online job marketplaces: towards a global labour market for outsourcing services? *Global Networks*, 15(2):236–255, 2015. https://doi.org/10.1111/glob.12051

13. John Horton, William R. Kerr, and Christopher Stanton. High-skilled migration to the united states and its economic consequences. In Gordon H. Hanson, William R. Kerr, and Sarah Turner, editors, *High-Skilled Migration to the United States and Its Economic Consequences*, chapter 3, pages 71–108. University of Chicago Press, 2018.

14. Phillip Brown, Hugh Lauder, and David Ashton. The global auction: The broken promises of education, jobs, and incomes. Oxford University Press, 2010.

15. Hjort Jonas and Poulsen Jonas. The arrival of fast internet and employment in africa. American Economic Review, 109(3):1032–79, 2019. https://doi.org/10.1257/aer.20161385

16. James Manyika, Armando Cabral, Lohini Moodley, Suraj Moraje, Safroadu Yeboah-Amankwah, Michael Chui, et al. Lions go digital: The Internet's transformative potential in Africa. Technical report, McKinsey Global Institute, 2013.

17. Kati Suominen. Fuelling Trade in the Digital Era. Issue Paper, International Centre for Trade and Sustainable Development, Geneva, 2017.

18. Vili Lehdonvirta, Helena Barnard, Mark Graham, and Isis Hjorth. Online labour markets-levelling the playing field for international service markets? Conference paper, Oxford Internet Institute, 2014.

19. Hong Yili and Pavlou Paul A. On buyer selection of service providers in online outsourcing platforms for it services. *Information Systems Research*, 28(3):547–562, 2017. https://doi.org/10.1287/isre.2017.0709

20. Ghani Ejaz, Kerr William R, and Stanton Christopher. Diasporas and outsourcing: evidence from odesk and india. *Management Science*, 60(7):1677–1697, 2014. https://doi.org/10.1287/mnsc.2013.1832

21. Ajay Agrawal, Nicola Lacetera, and Elizabeth Lyons. Does standardized information in online markets disproportionately benefit job applicants from less developed countries? *Journal of international Economics*, 103: 1–12, 2016. https://doi.org/10.1016/j.jinteco.2016.08.003

22. Balland Pierre-Alexandre, Cristian Jara-Figueroa, Petralia Sergio G, Steijn Mathieu PA, Rigby David L, and César A Hidalgo. Complex economic activities concentrate in large cities. *Nature Human Behaviour*, 4(3):248–254, 2020. https://doi.org/10.1038/s41562-019-0803-3 PMID:31932688

23. Fabian Braesemann, Vili Lehdonvirta, and Otto Kässi. Icts and the urban-rural divide: can online labour platforms bridge the gap? *Information, Communication & Society*, 25(1):34–54, 2022. https://doi.org/10.1080/1369118X.2020.1761857

24. Graham Mark and Anwar Mohammad Amir. The global gig economy: Towards a planetary labour market? *First Monday*, 24(4), 2019.

25. Mark Anner, Nicolas Pons-Vignon, and Uma Rani. For a future of work with dignity: A critique of the world 697 bank development report, the changing nature of work. *Global Labour Journal*, 10(1), 2019.

26. Mallett Richard W. Seeing the "changing nature of work" through a precarity len. *Global Labour Journal*, 11(3), 2020.

27. Wood Alex J, Graham Mark, Lehdonvirta Vili, and Hjorth Isis. Good gig, bad gig: Autonomy and algorithmic control in the global gig economy. *Work, Employment and Society*, 33(1):56–75, 2019. https://doi.org/10.1177/0950017018785616

28. Rani Uma and Furrer Marianne. On-demand digital economy: Can experience ensure work and income security for microtask workers? *Jahrbücher fur Nationalokonomie und Statistik*, 239(3):565–597, 2019. https://doi.org/10.1515/jbnst-2018-0019

29. Mark Graham, Isis Hjorth, and Vili Lehdonvirta. Digital labour and development: impacts of global digital labour platforms and the gig economy on worker livelihoods. *Transfer: European review of labour and research*, 23(2):135–162, 2017. https://doi.org/10.1177/1024258916687250

30. Anwar Mohammad Amir and Graham Mark. Hidden transcripts of the gig economy: labour agency and the new art of resistance among african gig workers. *Environment and Planning A: Economy and Space*, 52(7): 1269–1291, 2020. https://doi.org/10.1177/0308518X19894584

31. David H, Katz Lawrence F, and Kearney Melissa S. The polarization of the us labor market. *American economic review*, 96(2):189–194, 2006. https://doi.org/10.1257/000282806777212620

32. Kässi Otto and Lehdonvirta Vili. Do microcredentials help new workers enter the market? evidence from an online labor platform. *Journal of Human Resources*, pages 0519–10226R3, 2022.

33. Katharine A. Anderson. Skill networks and measures of complex human capital. *Proceedings of the National Academy of Sciences*, 2017.

34. Autor David H., Levy Frank, and Murnane Richard J. The skill content of recent technological change: An empirical exploration. *The Quarterly journal of economics*, 118(4):1279–1333, 2003. https://doi.org/10.1162/003355303322552801

35. Daron Acemoglu and David Autor. Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of labor economics*, volume 4, pages 1043–1171. Elsevier, 2011.

36. Maarten Goos, Alan Manning, and Anna Salomons. Explaining job polarization: Routine-biased technological change and offshoring. *American economic review*, 104(8):2509–26, 2014. https://doi.org/10.1257/aer.104.8.2509

37. Frey Carl Benedikt and Osborne Michael A. The future of employment: how susceptible are jobs to computerisation? *Technological forecasting and social change*, 114:254–280, 2017. https://doi.org/10.1016/j.techfore.2016.08.019

38. Blinder Alan S. How many US jobs might be offshorable? *World Economics*, 10(2):41, 2009.

39. David H and David Dorn. The growth of low-skill service jobs and the polarization of the us labor market. *American Economic Review*, 103(5):1553–97, 2013. https://doi.org/10.1257/aer.103.5.1553

40. Paul Collier. *The future of capitalism: Facing the new anxieties Penguin UK*, 2018.

41. Glasmeier Amy K. Income inequality and growing disparity: Spatial patterns of inequality and the case of the usa. *The new Oxford handbook of economic geography*, page 63, 2018.

42. Bettencourt Luís MA, Horacio Samaniego, and Hyejin Youn. Professional diversity and the productivity of cities. *Scientific reports*, 4(1):1–6, 2014.

43. Richard L Florida et al. *Cities and the creative class*. Psychology Press, 2005.

44. Boschma Ron A and Fritsch Michael. Creative class and regional growth: Empirical evidence from seven european countries. *Economic geography*, 85(4):391–423, 2009. https://doi.org/10.1111/j.1944-8287.2009.01048.x

45. Alan Benson, Aaron Sojourner, and Akhmed Umyarov. Can reputation discipline the gig economy? experimental evidence from an online labor market. *Management Science*, 66(5):1802–1825, 2020. https://doi.org/10.1287/mnsc.2019.3303

46. Horton John J. Buyer uncertainty about seller capacity: Causes, consequences, and a partial solution. *Management Science*, 65(8):3518–3540, 2019.

47. Barach Moshe A, Golden Joseph M, and Horton John J. Steering in online markets: the role of platform incentives and credibility. *Management Science*, 66(9):4047–4070, 2020. https://doi.org/10.1287/mnsc.2019.3412

48. Diane Perrons. Understanding social and spatial divisions in the new economy: New media clusters and the digital divide. *Economic geography*, 80(1):45–61, 2004.

49. Gordon Ian Richard and Kaplanis Ioannis. Accounting for big-city growth in low-paid occupations: Immigration and/or service-class consumption. *Economic Geography*, 90(1):67–90, 2014. https://doi.org/10.1111/ecge.12026

50. Chris Forman, Avi Goldfarb, and Shane Greenstein. How Geography Shapes—and is Shapey by—the Internet. In *The New Oxford Handbook of Economic Geography*, pages 269–285. Oxford University Press, 2018.

51. Mark Graham. Warped geographies of development: The Internet and theories of economic development. *Geography compass*, 2(3):771–789, 2008. https://doi.org/10.1111/j.1749-8198.2008.00093.x

52. Henry Wai-Chung Yeung. The Logic of Production Networks. In *The New Oxford Handbook of Economic Geography*, pages 382–406. Oxford University Press, 2018.

53. Bauer Johannes M. The Internet and income inequality: Socio-economic challenges in a hyperconnected society. *Telecommunications Policy*, 42(4):333–343, 2018. https://doi.org/10.1016/j.telpol.2017.05.009

54. Blinder Alan S. and Krueger Alan B. Alternative measures of offshorability: a survey approach. *Journal of Labor Economics*, 31(S1):S97–S128, 2013. https://doi.org/10.1086/669061

55. Stephan Manning, Marcus M. Larsen, and Chacko G. Kannothra. Global sourcing of business processes: History, effects, and future trends. The new Oxford handbook of economic geography, pages 407–435, 2017. Publisher: Oxford University Press Oxford.

56. Melanie Feakins. Off-offshoring from russia to ukraine: How russian transnational entrepreneurs created a post-soviet it offshore. *Economic Geography*, 93(4):397–417, 2017. https://doi.org/10.1080/00130095.2017.1308222

57. Malecki Edward J. The economic geography of the internet's infrastructure. *Economic geography*, 78(4): 399–424, 2002. https://doi.org/10.2307/4140796

58. Vu Khuong, Hanafizadeh Payam, and Bohlin Erik. Ict as a driver of economic growth: A survey of the literature and directions for future research. *Telecommunications Policy*, 44(2):101922, 2020. https://doi.org/10.1016/j.telpol.2020.101922

59. Carol Corrado and Bart van Ark. The internet and productivity. In *Handbook on the Economics of the Internet*. Edward Elgar Publishing, 2016.

60. Bettencourt Luís MA, Horacio Samaniego, and Hyejin Youn. Professional diversity and the productivity of cities. *Scientific reports*, 4:5393, 2014.

61. Roberto Gallardo, Brian Whitacre, Indraneel Kumar, and Sreedhar Upendram. Broadband metrics and job productivity: a look at county-level data. *The Annals of Regional Science*, 66(1):161–184, 2021. https://doi.org/10.1007/s00168-020-01015-0

62. Dario Diodato, Frank Neffke, and Neave O'Clery. Why do industries coagglomerate? how marshallian externalities differ by industry and have evolved over time. *Journal of Urban Economics*, 106:1–26, 2018.

63. Ajay Agrawal, John Horton, Nicola Lacetera, and Elizabeth Lyons. Digitization and the contract labor market: A research agenda. In *Economic analysis of the digital economy*, pages 219–250. University of Chicago Press, 2015.

64. Lehdonvirta Vili and Corporaal Greetje. Platform Sourcing: How Fortune 500 Firms are Adopting Online Freelancing Platforms. *Oxford Internet Institute*, 2017.

65. Kathrin Borchert, Matthias Hirth, Michael E Kummer, Ulrich Laitenberger, Olga Slivko, and Steffen Viete. Unemployment and online labor. *ZEW-Centre for European Economic Research Discussion Paper*, (18–023), 2018.

66. Lukac Martin and Grow Andre. Reputation systems and recruitment in online labor markets: insights from an agent-based model. *Journal of Computational Social Science*, 4(1):207–229, 2021. https://doi.org/10.1007/s42001-020-00072-x

67. Amanda Pallais. Inefficient hiring in entry-level labor markets. *American Economic Review*, 104(11):3565–99, 2014. https://doi.org/10.1257/aer.104.11.3565

68. Stanton Christopher T and Thomas Catherine. Landing the first job: The value of intermediaries in online hiring. *The Review of Economic Studies*, 83(2):810–854, 2016. https://doi.org/10.1093/restud/rdv042

69. Jeroen Smits. GDL Area Database. GDL Working Paper 16-101, Global Data Lab, Institute for Management Research, Radboud University, 2016.

70. Burbidge John B., Magee Lonnie, and Robb A. Leslie. Alternative Transformations to Handle Extreme Values of the Dependent Variable. *Journal of the American Statistical Association*, 83(401):123–127, 1988. https://doi.org/10.1080/01621459.1988.10478575

71. Emil Mihaylov and Kea Gartje Tijdens. Measuring the routine and non-routine task content of 427 four-digit isco-08 occupations. Working paper, SSRN, 2019.

72. Pallais Amanda and Sands Emily Glassberg. Why the referential treatment? Evidence from field experiments on referrals. *Journal of Political Economy*, 124(6):1793–1828, 2016. https://doi.org/10.1086/688850

73. Adam Ozimek. The future of remote work. *Available at SSRN 3638597*, 2020.

74. Jose Maria Barrero, Nicholas Bloom, and Steven J Davis. Why working from home will stick. Technical report, National Bureau of Economic Research, 2021.

75. Behrendt Christina, Nguyen Quynh Anh, and Rani Uma. Social protection systems and the future of work: Ensuring social security for digital platform workers. *International Social Security Review*, 72(3):17–41, 2019. https://doi.org/10.1111/issr.12212

76. Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, et al. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1301–1318, 2013.

77. Eichhorst Werner, Hinte Holger, Rinne Ulf, and Tobsch Verena. How big is the gig? assessing the preliminary evidence on the effects of digitalization on the labor market. *Management revue*, pages 298–318, 2017.

78. Panagiotis G Ipeirotis and John J Horton. The need for standardization in crowdsourcing. In *Proceedings of the workshop on crowdsourcing and human computation at CHI*, 2011.

79. Mika Pajarinen, Petri Rouvinen, JOrg Claussen, Jari Hakanen, Anne Kovalainen, Tobias Kretschmer, et al. Upworkers in finland: Survey results. Technical report, ETLA Report, 2018.

80. Young Alwyn. Inequality, the urban-rural gap, and migration. *The Quarterly Journal of Economics*, 128(4): 1727–1785, 2013. https://doi.org/10.1093/qje/qjt025

81. Martin Lukac. Two worlds of online labour markets: Exploring segmentation using finite mixture models and a network of skill co-occurrence. Working paper, SocArXiv, 2021.

82. Lin Mingfeng, Liu Yong, and Viswanathan Siva. Effectiveness of reputation in contracting for customized production: Evidence from online labor markets. *Management Science*, 64(1):345–359, 2018. https://doi.org/10.1287/mnsc.2016.2594

83. Stark David and Pais Ivana. Algorithmic management in the platform economy. *Sociologica*, 14(3):47–72, 2020.

84. Mareike MOhlmann, Lior Zalmanson, Ola Henfridsson, and Gregory Robert Wayne. Algorithmic management of work on online labor platforms: When matching meets control. *MIS Quarterly*, 45(4), 2021.

85. Sveikauskas Leo. The productivity of cities. *The Quarterly Journal of Economics*, 89(3):393–413, 1975. https://doi.org/10.2307/1885259

86. Quigley John M. Urban diversity and economic growth. *Journal of Economic perspectives*, 12(2):127–138, 1998. https://doi.org/10.1257/jep.12.2.127

87. Bettencourt Luís MA. The origins of scaling in cities. *science*, 340(6139):1438–1441, 2013. https://doi.org/10.1126/science.1235823

88. Bettencourt Luís MA, Lobo Jose, Helbing Dirk, Kuhnert Christian, and West Geoffrey B. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences*, 104(17):7301–7306, 2007. https://doi.org/10.1073/pnas.0610172104

89. Bettencourt Luís MA, Lobo Jose, Strumsky Deborah, and West Geoffrey B. Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. *PloS one*, 5(11):e13541, 2010. https://doi.org/10.1371/journal.pone.0013541

2.9 Case Study 3

RESEARCH ARTICLE

Automation, workers' skills and job satisfaction

Henrik Schwabe, Fulvio Castellacci

TIK Centre, University of Oslo, Oslo, Norway

Abstract

When industrial robots are adopted by firms in a local labor market, some workers are displaced and become unemployed. Other workers that are not directly affected by automation may however fear that these new technologies might replace their working tasks in the future. This fear of a possible future replacement is important because it negatively affects workers' job satisfaction at present. This paper studies the extent to which automation affects workers' job satisfaction, and whether this effect differs for high- versus low-skilled workers. The empirical analysis uses microdata for several thousand workers in Norway from the *Working Life Barometer* survey for the period 2016–2019, combined with information on the introduction of industrial robots in Norway from the International Federation of Robotics. Our identification strategy exploits variation in the pace of introduction of industrial robots in Norwegian regions and industries since 2007 to instrument workers' fear of replacement. The results indicate that automation in industrial firms in recent years have induced 40% of the workers that are currently in employment to fear that their work might be replaced by a smart machine in the future. Such fear of future replacement does negatively affect workers' job satisfaction at present. This negative effect is driven by low-skilled workers, which are those carrying out routine-based tasks, and who are therefore more exposed to the risks of automation.

1. Introduction

Industrial robotics and artificial intelligence (AI) have in the last few years increasingly been used in production activities. This has led to the automation of many tasks that were previously carried out by workers, and that can now be performed by smart machines. The fear that these technological advances may have dramatic consequences on the future of labor has fostered the recent development of new economics research studying the effects of automation on employment [1,2]. Recent models and empirical evidence on this topic show that automation can have negative effects on employment demand and wages, and particularly so for workers that perform routine-based tasks that can more easily be displaced [3,4]. On the other hand, however, these new technologies may also have positive effects by increasing productivity [5].

This recent research has so far focused on the effects of automation, industrial robots and artificial intelligence on labor demand and wages. However, while employment and wages are two central dimensions shaping individual workers' well-being, it is also important to point out that other non-pecuniary aspects do contribute to shape workers' well-being, and that automation may potentially have important impacts on these [6]. Specifically, if workers fear that their occupation might be replaced by a smart machine in the future, such prospect and uncertainty about future working conditions may arguably affect their job satisfaction at present [7,8].

Why should we care about the impacts of automation on workers' job satisfaction? The reason is twofold. First, since individuals spend a substantial part of their life at work, job satisfaction experienced in working

https://doi.org/10.1371/journal.pone.0242929 Used with CC 4.0 license

life does indeed represent an important component of individuals' overall subjective well-being [9]. Second, workers that are not happy and experience dissatisfaction with their job have typically lower motivation and efforts [10], and higher turnover rates. Therefore, if a large number of workers in the economy fear to be replaced by smart machines in the future, this fear may lead to mental stress and anxiety at present, as well as hamper productivity and innovation in the economy.

In spite of the relevance of this topic, to the best of our knowledge only two papers have recently explored the relationship between automation and workers' well-being. Abeliansky and Beulmann [11] focuses on workers' mental health in Germany; and Schwabe [12] studies workers' life satisfaction in a sample of European countries. Neither of these studies, though, investigates explicitly the impacts of automation on job satisfaction.

Further, these recent works do not study the role of workers' skills, and how these may affect the relationship between automation and well-being. The literature on automation and employment clearly shows that the effects of the introduction of industrial robots largely differ for high-skilled and low-skilled workers. It is therefore paramount to investigate whether the effects of automation on job satisfaction can have different effects on workers' well-being depending on their skill levels. In short, the question investigated in the present paper is the following: *Does automation affect workers' job satisfaction–and how does this effect differ for high- versus low-skilled workers*?

To study this question, it is useful to distinguish two related dimensions. The first side of the link between automation and job satisfaction is that the introduction of industrial robots in local labor markets will affect workers' expectations about their future jobs, i.e. it will lead some workers to fear that part of their working tasks might be replaced by a smart machine in the future. The second dimension is that these expectations about the future, and particularly the anticipated fear of replacement, will negatively affect workers' subjective well-being at present.

Empirically, we operationalize this idea by making use of a two-stage econometric model, in which fear of replacement and job satisfaction are the dependent variables of the first and the second stage, respectively. The empirical analysis uses microdata for several thousand workers in Norway from the *Working Life Barometer* survey (*Arbeidslivsbarometer*) (four annual surveys for the period 2016–2019), combined with information on the introduction of industrial robots in Norway from the *International Federation of Robotics* (IFR) dataset. Our identification strategy exploits variation in the pace of introduction of industrial robots in Norwegian regions and industries between 2007 and t (i.e. the time at which each of the four surveys was carried out).

The results indicate that automation in industrial firms in recent years has induced workers that are currently in employment to fear that their work might be replaced by a smart machine in the future, and that this effect is stronger for low-skilled workers. Further, our findings show that fear of future replacement does negatively affect workers' job satisfaction at present, and that such negative effect is in particular significant for low-skilled workers, which are those carrying out routine-based tasks, and who are therefore more exposed to the risks of automation.

On the whole, these results contribute to, and extend, the recent literature on automation and employment by shifting the focus to important nonpecuniary impacts that are reflected in workers' expectations, fears and job satisfaction, and showing that workers' skills is an important variable moderating the effects of automation on subjective well-being.

The paper is organized as followed. Section 2 reviews the literature on automation and employment. Section 3 points out the conceptual mechanisms that are relevant to explain the effects of automation on job satisfaction.

https://doi.org/10.1371/journal.pone.0242929

Section 4 presents the data and indicators. Section 5 discusses the empirical methods. Section 6 presents the results. Section 7 concludes and discusses the main contributions and implications.

2. Literature

Effects of Automation on Employment and Wages

Automation, industrial robotics and artificial intelligence have in the last few years experienced substantial advances and found an increasing number of applications in production activities. Artificial intelligence and robotics have developed as two distinct scientific and technological fields for a long time, and only recently they have converged and cross-fertilized [13]. Frank et al. [2] presents relevant illustrations of this recent convergence, and it discusses challenges for research on the economic effects of AI and automation. This has spurred the recent development of a strand of scholarly research studying the effects of these new technologies on employment.

A starting point of this literature is the canonical model of skilled bias [14], according to which new skilled-bias technologies lead to polarization and increasing differences in employment opportunities and wages between skilled and unskilled workers. Sachs and Kotlikoff [15] present a simple framework in which smart machines substitute directly for young unskilled labor, whereas they are complementary to older skilled workers. Young unskilled workers experience lower wages, which in turn lead to lower saving and investments in human and physical capital–thus perpetuating and strengthening the gap between young unskilled and older skilled workers over time.

Such pessimistic prediction on the future of employment is however not shared by other works in this field. Taking a long-run historical perspective, Autor [16] and Mokyr et al. [1] argue that, as in other times in history, technological progress will lead to major structural changes in the quantity and content of work, but it will arguably not lead to a complete substitution of capital for labor. Houseman [17] provides empirical evidence that, although manufacturing employment in the US has declined since early 2000s, this is mainly explained by international trade and global competition effects, and there is weak support in the data for the argument that such decrease in employment is due to automation. More recently, McGuinness et al. [18] and Klenert et al. [19] present empirical studies that indicate that automation technologies and industrial robots have actually positive effects on employment. On the one hand, automation leads to a creative destruction process that may on the whole increase the overall demand for labor. On the other hand, it may contribute to reduce routine-based working tasks, which are typically monotonous and physically straining, thus improving the quality of work.

A more nuanced perspective that considers both negative and positive effects of automation on employment is presented by studies of the job polarization hypothesis. In short, the main idea of this research is that automation technologies complement highly skilled labor, explaining its expansion and wage growth in recent years in most advanced countries. On the other hand, middle-skilled workers are those more negatively affected by routine-biased technical change, because their tasks are relatively easier to automate. As for low-skilled workers, and particularly those employed in personal services occupations, these often perform manual and personal communication tasks that are not that easy to automate yet. Hence, the resulting pattern is that middle-skilled workers have in recent years shifted towards low-skilled employment occupations, which have consequently grown and experienced higher wages. All in all, this explains the observed increasing polarization in the job market, with the growth of employment and wages for high- and low-skilled workers, and a corresponding decline for middle-skilled occupations [3,4,16,20]. Beaudry et al. [21] argue however that the demand for high-skilled workers has declined after 2000 due to decreasing returns to investments in information and communication technologies (ICTs), and that high-skilled have then begun to compete for

lower-skilled jobs. This study, though, is based on empirical evidence on ICT investments in general, and it does not focus specifically on the effects of AI and automation.

Acemoglu and Restrepo [22] present a theoretical framework that is useful to study both negative and positive effects of industrial robots on employment and wages. The model points out two contrasting effects of industrial automation: a *displacement* effect that negatively affects the demand for employment and the wages of workers that perform routine-based tasks; and a *productivity* effect that creates benefits for workers that perform non-routine tasks (in the automated sector as well as in other sectors and occupations of the economy). This study also presents empirical evidence that corroborates the model's predictions on the effects of industrial robots on employment and wages in US manufacturing industries between 1990 and 2007. In line with evidence presented by other recent works [5,23,24], their results show that overall the displacement (negative) effect of the introduction of industrial robots has until now been stronger than the productivity (positive) effect.

Effects of Automation on Job Satisfaction

This recent strand of research has so far focused on the effects of automation, industrial robots and artificial intelligence on aggregate patterns of labor demand and wages for different countries and industries. However, research has not investigated yet the impacts that these new technologies may have on individual workers' subjective well-being. Do workers fear that their occupation might be replaced by a smart machine in the future, and if so how does that prospect affect their current job satisfaction?.

Job satisfaction is the subjective well-being of workers (i.e. their own assessment of the well-being they experience at work). This is an obviously crucial dimension for economic analysis and policy. First, since individuals spend a substantial part of their life at work, job satisfaction experienced in working life represents an important component of individuals' overall subjective well-being. Second, workers that are not happy and experience dissatisfaction with their job have typically lower motivation and efforts, and higher turnover rates. This, in turn, weakens productivity and innovation in the economy.

The literature on job satisfaction is wide-ranging, and it has extensively investigated a variety of factors that explain why some individuals report higher subjective well-being than others [7,8]. However, only a few studies have so far explicitly investigated the relationships between the widespread diffusion and application of digital technologies and job satisfaction [25]. Kaplan and Schulhofer-Wohl[6], using data from the American Time Use survey, discusses the nonpecuniary implications of changes in the occupational structure in the US in recent decades, i.e. the effects of these structural changes on different aspects of job satisfaction such as reported happiness, stress and meaning at work. The work indicates that the changing occupational structure has not only led to polarization in terms of skills and wages, but it has also determined substantial changes in workers' feelings about the job they have and the tasks they perform.

Two recent papers explore the relationship between automation and workers' well-being. Abeliansky and Beulmann [11] present an empirical study on the impact of automation on the mental health of workers (which is one important dimension reflecting stress and weak job satisfaction). The analysis uses individual-level data from the German Socioeconomic Panel for the period 2002–2014 linked to industry-level data on use of industrial robots in 21 manufacturing sectors in Germany. The results indicate that automation negatively affects workers' mental health, and this effect is related to the fear of having lower wages and worse economic conditions in the future.

Schwabe [12] makes use of worker-level data from the Eurobarometer survey for European countries (period 2012–2017) to investigate the relationships between fear of replacement and workers' subjective well-being (measured by life satisfaction, which is as well-known an evaluative dimension of individuals' well-being).

The results of this study find that fear of replacement affects life satisfaction, but the direction of this effect does largely depend on age. In line with models of skill-bias and job polarization (see section 2.1), younger workers regard replacement as a possible threat to their job opportunities in the future, whereas older workers look at it as a positive technological development that is not likely to affect them directly, and that will arguably enhance well-being and prosperity in the society.

These two studies provide an important starting point for the present work. None of them, though, investigates explicitly the role of workers' skills, which is however a key dimension in the literature on the employment effects of automation briefly reviewed in section 2.1. In the job satisfaction literature too, education and skill levels represent one of the central factors affecting the job satisfaction of workers [26].

Two contrasting mechanisms link education and job satisfaction. On the one hand, a higher skill level increases the chances that an employee will have a higher wage level and a more interesting and rewarding job, which enhance job satisfaction. On the other hand, however, various empirical studies have found that–after controlling for income earnings–the correlation between education level and subjective well-being at work is negative [8,27,28]. This can be explained in the light of prospect theory [29]. When an individual invests more time in education and human capital formation, her expectations about the desired job will also be higher, and it will therefore be more likely that the worker will feel more critical and less satisfied with her actual working conditions if these high expectations are unmet. In particular, empirical research indicates that overqualified workers report significant lower levels of job satisfaction than others [26,30].

3. Question and Propositions

The question investigated in the present paper is the following: *Does automation affect workers' job satisfaction–and how does this effect differ for high- versus low-skilled workers*? The first part of the question refers to the main impact of automation on job satisfaction, which as noted above has not been analyzed in previous research yet. The second part of the question suggests that fear of replacement can have different effects on workers' well-being depending on their skill levels, and it seeks to investigate these moderation effects.

Conceptually, the link between automation and job satisfaction can be analyzed in two steps. The first is that the introduction of industrial robots in local labor markets will arguably affect workers' expectations about their future jobs, which means that some workers will fear that some of their tasks, or even their whole job, might be replaced by a smart machine in the future. The second step is that these expectations about the future, and particularly the anticipated fear of replacement, will affect workers' job satisfaction at present.

Our empirical analysis will consider both of these conceptual steps in a two-stage empirical model, and investigate whether the related impacts are stronger for high-skilled or for low-skilled workers. We point out below here the main effects that we expect to find in the empirical analysis, and how these can be explained in the light of the literature reviewed in this section. As noted below, some of the effects of interest are stronger for high-skilled workers, whereas others are more relevant for low-skilled workers, so that the overall net moderation effect cannot be pointed out *ex-ante*, but it will have to be established based on the empirical evidence.

I. Fear of Replacement

The introduction of industrial robots in the local labor market increases the likelihood that some workers will be replaced by smart machines in the future. These technological changes and their applications in firms in local labor markets will therefore induce some workers that are currently employed to fear that they might be replaced in the future (or at least that some of their tasks might be).

Moderation effects. The introduction of industrial robots will arguably have different impacts for high- versus low-skilled workers. We envisage two contrasting effects.

Fear of replacement is stronger for the low-skilled. These workers are more exposed to the risks of displacement from automation because they typically carry out routine tasks that can more easily be automated (see literature in section 2.1).

Fear of replacement is stronger for the high-skilled. High-skilled workers are typically also more educated individuals who read more and follow media debates on robots, automation and their negative consequences for employment. Hence, high skilled workers are arguably more exposed to peer effects, which may translate in a greater fear about the future of employment. Contrary to this argument, we may however posit that workers of higher education typically have a better ability to understand and anticipate that these new technologies will also have positive effects for their future tasks and wages, as well as for the overall productivity of the economy–i.e. they are arguably be more forward-looking [31].

Proposition 1: The introduction of industrial robots in the local labor market will negatively affect low-skilled workers more than high-skilled workers if the former effect is stronger than the latter.

II. Job Satisfaction

The second aspect of our conceptual analysis refers to the impacts that fear of replacement will have for workers' subjective well-being. The main expectation is that fear of replacement in the future will negatively affect job satisfaction at present. The main reason is that the prospect to become unemployed, or to be taken away some of the current working tasks, will negatively affect wage and financial conditions expected for the future, thus creating uncertainty about future job prospects and personal finance, and hence lower job satisfaction.

Moderation effects. Fear of replacement will arguably have different impacts on job satisfaction for high- versus low-skilled workers. We posit the following contrasting effects.

The negative effects on job satisfaction will be stronger for the low-skilled. If replaced, these workers will on average have fewer possibilities to find another occupation in the labor market. Acemoglu and Restrepo [22] and Blanas et al. [20] document in fact that displacement effects of industrial robots on employment and wages are stronger and more significant for low-education workers. On the other hand, as noted in section 2.1, extant research suggests that automation technologies can have more positive effects on high-skilled workers, increasing the demand for labor, wages and the complexity and interest of their tasks [18].

The negative effects on job satisfaction will be stronger for the high-skilled. According to prospect theory [29], individuals that invest more time in education and human capital formation will also have higher expectations about the working conditions that they desire and expect to have in the future, and be less satisfied with their job if this does not match the high expectations the individual has. Hence, highly educated workers, when facing the prospect of changing jobs and tasks in the future, may be those that have more to lose from automation, precisely because they are the individuals who have invested more in their human capital formation, and they have therefore higher expectations about the working conditions that they feel they deserve.

Proposition 2: Fear of replacement will negatively affect the job satisfaction of low-skilled workers more than that of high-skilled workers if the former effect is stronger than the latter.

https://doi.org/10.1371/journal.pone.0242929

4. Data

Individual-Level Data

We use the *Working Life Barometer* survey (*Arbeidslivsbarometer*), which provides annual microdata for several thousand Norwegian workers. The survey is provided by the Confederation of Vocational Unions (YS), a politically independent umbrella organization for labor unions, and organized by the Work Research Institute in Norway. TNS Gallup collects the data targeting a large random sample of Norwegian workers aged 18–67 years. Our analysis makes use of the four surveys carried out in the years 2016 to 2019, which include information on the main variables of interest for this study, and particularly workers' subjective assessments of the threats of automation, and their job satisfaction.

The main target variable in the study is job satisfaction, which is measured by means of responses to the survey question: "*How satisfied are you with your job*?." Respondents indicate their satisfaction level on a 1–5 scale ("Very dissatisfied"; "Pretty dissatisfied"; "Neither satisfied nor dissatisfied"; "Pretty satisfied"; "Very satisfied"). The main explanatory variable is fear of replacement. This is measured by means of responses to the following survey question: "*Do you think some of your current tasks could be done by machine instead*?." Fear of replacement is a dummy variable: respondents who answer yes to this question take value 1, whereas workers who do not think that their tasks could be replaced by a machine take value 0. It is important to observe that this survey question measures workers' assessment of the possibility that their tasks could be replaced by machines (cognitive reaction), and not directly the fear to lose their job as a consequence of automation (emotional reaction). However, as we will show later in the results section, this survey question is closely related to other survey questions that measure workers' fear of losing their job, and it is therefore reasonable to use it as a proxy measure of fear of replacement. It is also worthwhile to note that only workers who are currently employed are asked to answer the question on fear of replacement, whereas unemployed individuals must skip this part of the questionnaire. Hence, our analysis focuses on the beliefs of workers who are potentially exposed to automation, but it does not consider those individuals that have already been laid off due to automation.

Next, another important variable in our study is the skill-level of workers, which is measured by their education level, distinguishing workers with a completed University degree *versus* those without tertiary education. In terms of control variables, the *Working Life Barometer* survey also provides employee-level demographic and socio-economic information such as age, gender, income, union membership, and occupation type. In total, we analyze responses from 10,051 workers aged 19–68 years. Table 1 presents a list of the variables used in the analysis, and Table 2 reports some descriptive statistics.

Table 2.5 Variables.

Variable	Definition
	Individual level variables
Job satisfaction	Respondents indicate their job satisfaction ranging from 1 "Very dissatisfied"; 2 "Pretty dissatisfied"; 3 "Neither satisfied nor dissatisfied"; 4 "Pretty satisfied"; 5 "Very satisfied".
Machine replacement	Respondents indicate whether they believe that a machine can perform some of their job tasks.
Union membership	Dummy indicating whether the respondent is unionized.
Age	Age of respondent.
Women	Dummy indicating the gender of the respondent.
University degree	Dummy indicating whether the respondent has a university degree.

https://doi.org/10.1371/journal.pone.0242929

Variable	Definition
Working in industry	Dummy indicating whether the respondent is an industry worker.
	Regional level variables
ΔRobot exposure	Industry-region's long-term robot adoption per thousand workers. More detailed definition in main text.
Unemployment benefit recipients	Share of regional population that are registered recipients of unemployment benefits.
Business building broadband infrastructure availability	Fixed broadband penetration per 100 inhabitants.
Population	Log of regional population.
GDP per capita	Log of regional GDP per capita.
Tertiary education	Regional share of population (aged 25–64) with tertiary education.
Share of big industrial companies	Big industrial companies as share of total firm population by region.

https://doi.org/10.1371/journal.pone.0242929.t001

Robot Data

To measure the introduction of industrial robots in local labor markets in Norway, we make use of a dataset provided by the International Federation of Robotics (IFR), which contains information on robot stock and deliveries in Norwegian industrial firms since 1993. The IFR defines an industrial robot as an "*automatically controlled, reprogrammable multipurpose [stationary or mobile machine]*" [32]. Following this definition, industrial robots are autonomous machines capable of operating without human intervention and that could potentially substitute or complement human labor. The IFR provides detailed data on robot stock and deliveries for the period 1993–2017, which can be broken down by application or industry. Robot stock for years 2018 and 2019 are extrapolated assuming a 9 percent annual growth in operational stock as projected by IFR [33]. IFR data have recently been used to analyze the impact of automation on employment and wages [22,34,35], as well as on workers' well-being [11,12].

Table 2.6 Descriptive statistics.

Variable	Obs.	Mean	Std. Dev.	Min	Max
Job satisfaction*	10,051	3.99	0.84	1.00	5.00
Machine replacement	10,051	0.40	0.49	0.00	1.00
Union membership	10,051	0.69	0.46	0.00	1.00
Age	10,051	46.44	11.67	19.00	68.00
Income scale	10,051	4.81	1.81	1.00	9.00
Women	10,051	0.52	0.50	0.00	1.00
University degree	10,051	0.56	0.50	0.00	1.00
Working in industry	10,051	0.08	0.26	0.00	1.00
Robot exposure	10,051	0.06	0.03	0.00	0.20
Log(GDP per capita)	10,051	12.89	0.53	12.14	13.69
Tertiary education (share of population)	10,051	43.81	6.68	35.5	54.3

(*Continued*)

https://doi.org/10.1371/journal.pone.0242929

Table 2.6 Descriptive statistics. (*Continued*)

Variable	Obs.	Mean	Std. Dev.	Min	Max
Business building broadband infrastructure availability	10,051	0.74	0.14	0.56	0.97
Log(population)	10,051	14.09	0.20	13.74	14.33
Unemployment benefit recipients (share of population)	10,051	4.34	0.45	3.41	5.12
Share of big industrial companies	10,051	10.72	1.20	8.05	12.61

* Job satisfaction: Very dissatisfied: 1.17%; pretty dissatisfied: 4.98%; neither satisfied nor dissatisfied: 13.57%; pretty satisfied: 53.86%: Very satisfied: 26.42%.

https://doi.org/10.1371/journal.pone.0242929.t002

We allocate robots in regional labor markets following extant research [22,34,36], assuming that robots are distributed across region and industries by their respective employment shares. Employment shares are calculated based on Eurostat's Labor Force Survey data dating back to 2008. The long-term change in robot adoption occurs between years 2008 and t based on initial regional employment composition in each industrial category (industry, agriculture, construction, and services), with the change in robot adoption per 1,000 workers fixed at the starting level in year 2008.

$$\Delta robot\ exposure_{r,s,t} = \sum_{s \in S} \frac{emp_{r,s,2008}}{emp_{r,2008}} * \left(\frac{robots_{s,t} - robots_{s,2008}}{emp_{c,2007}} \right) \quad (1)$$

In this setup, robot exposure is measured as national robot adoption allocated at the region-industry level (r,s). Each regional labor market r is scaled by the nation's total employment emp_c. In short, the instrumental variable that we will use in our empirical analysis is the long-term change in the adoption of industrial robots by Norwegian firms in each local labor market (i.e. in each region-industry r,s). This measures the extent to which workers have been exposed to automation from 2007 onwards (see a further discussion of the empirical identification strategy in section 3.2 below).

Regional-Level Control Variables

We use the Eurostat's Labor Force Survey to obtain regional-level variables on GDP per capita, population share with tertiary education, and population size. From Statistics Norway, we retrieve data on firms by size for each region. Further, we collect data on unemployment benefit recipients as a share of total population from the Norwegian Labour and Welfare Administration (NAV), for each region and each year of our dataset.

To avoid omitting the possible conflating influence of ICTs when analyzing automation, previous studies have included ICT capital or investment as an additional control variable [34,37]. However, others argue that more specific measures of ICT utilization are necessary for micro-level studies [38]. Unlike existing studies that have analyzed the impact of high-speed broadband developments in Norway [39,40], we use as additional control variable the broadband internet availability in office buildings instead of households in each region. Data on office buildings with at least 8/8 Mbit/s speeds are provided by the Norwegian Communications Authority (Nkom), and matched against individuals through regional identifiers.

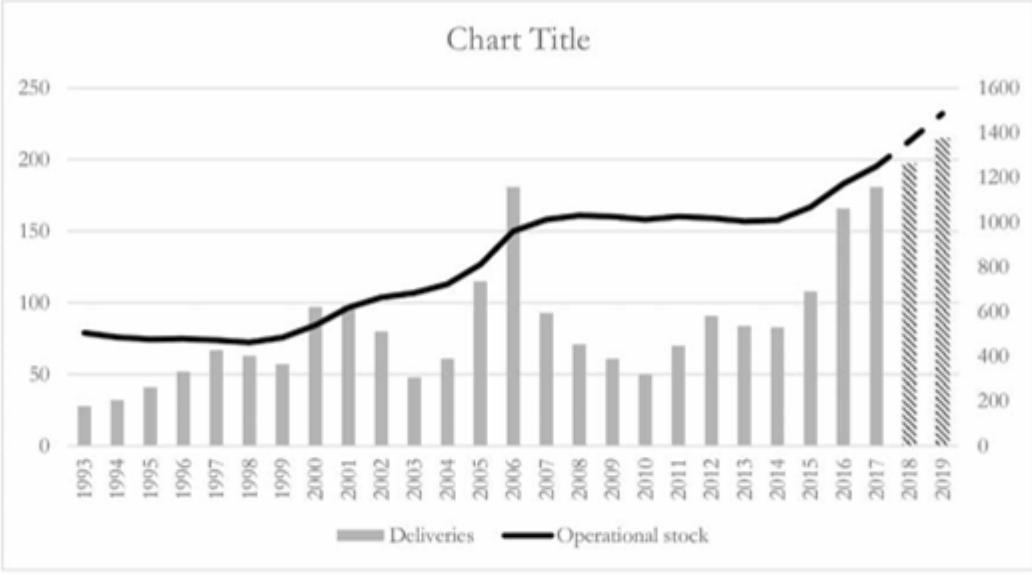
5. Empirical Methods

The econometric analysis sets out to study the relationship between fear of replacement and job satisfaction. Fear of replacement is the subjective assessment that each worker does on the possibility that her working tasks will be replaced by a smart machine in the future. Such subjective assessment may arguably depend on unobserved and idiosyncratic factors such as e.g. ability, attitude towards risk, and technological/digital competencies. Therefore, unobserved individual factors might possibly influence both the outcome variable (job satisfaction) and the main explanatory variable (fear of replacement).

To address endogeneity concerns, we follow recent research and use the lagged introduction of robots in local labor markets (industry-regions) as an instrument for individual workers' fear of replacement [11,12]. Existing studies on robot implications for labor markets where robot adoption is the main explanatory variable address endogeneity issues by incorporating spillover effects from robot adoption across industries in other countries as an instrument in a 2SLS setup [22,34,36]. Unlike these studies, we approach subjective responses to structural inroads of robot technology in local labor markets to identify learning effects from past automation. Specifically, our instrumental variable is the one defined in (1) above, i.e. the change in the adoption of industrial robots by Norwegian firms in each local labor market (industry-region) between 2008 and year t (i.e. one of the survey years 2016–2019). This variable measures the extent to which workers in each of the 16 industry-regions considered in this study have been *exposed* to rising automation in recent years. We thus exploit (lagged) variation in robot adoption over time and across industry-regions in Norway to instrument for individual fear of replacement at time t. The underlying idea of this identification strategy is that workers that are employed in local labor markets that have more rapidly been exposed to automation (i.e. in industry-regions where firms have increasingly used industrial robots) will be more likely to consider automation as a possible threat, and therefore fear that some of their working tasks could be replaced by a machine in the future. In other words, we posit that workers learn from past robot adoption in their local labor markets, because they are subject to *peer effects* [41]. Although it is reasonable to posit that these peer effects work through *changes* in robot adoption over time, we cannot exclude the possibility that the same mechanism may also work through the absolute *levels* of robot adoption (i.e. workers may fear replacement when they experience a high intensity of industrial robots in the industry-region where they work). To consider this possibility, we have also calculated our instrumental variable in levels rather than as changes over time, and reported additional regressions in the online appendix (see Table A5 in S1 File, whose results are in line with the main results presented in the paper).

Norwegian firms have invested in sophisticated robotics and automation technologies to keep pace with the Digital Single Market strategy [42], and our empirical analysis exploits this exogenous source of tempo-spatial variations to identify the effects of automation on workers' job satisfaction. Fig 1 illustrates the dynamics of industrial robots adoption in Norway in the last decades, showing a much faster pace since 2014. Table 3 shows that most robots have so far been used by firms within manufacturing, and less so in other branches such as agriculture, construction and services. However, Table 3 also shows that the introduction of robots by service firms has been quite rapid in the last decade. Fig 2 illustrates the trend in robot adoption since 2010, indicating a rising trend in all 16 industry-regions considered in this study, and particularly so in manufacturing and services.

https://doi.org/10.1371/journal.pone.0242929



The data for 2018 and 2019 are estimated (see data section).

Fig 2.4 Robot deliveries and operational stock for Norway between 1993 and 2019. The data for 2018 and 2019 are estimated (see data section).

https://doi.org/10.1371/journal.pone.0242929.g001

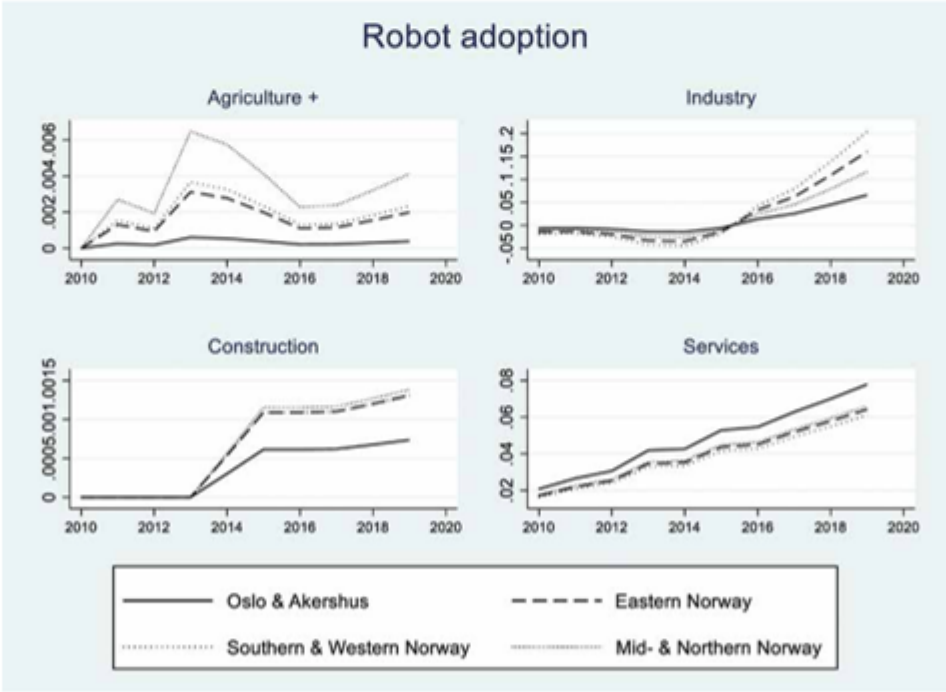


Fig 2.5 Robot adoption by industry-region, 2010–2019.

https://doi.org/10.1371/journal.pone.0242929.g002

Table 2.7 Adoption of robots (operational stock) in Norwegian regions and industries.

Region	Sector	2008	2017
Oslo & Akershus	Agriculture, forestry and fishing	0	1
	Industry	113	140
	Construction	0	0
	Services	3	13
Eastern Norway	Agriculture, forestry and fishing	3	4
	Industry	296	294
	Construction	0	1
	Services	3	11
Southern & Western Norway	Agriculture, forestry and fishing	3	4
	Industry	444	521
	Construction	0	1
	Services	3	14
Mid- and Northern Norway	Agriculture, forestry and fishing	3	5
	Industry	141	173
	Construction	0	0
	Services	3	13

https://doi.org/10.1371/journal.pone.0242929.t003

To get a further overview of the diffusion and use of industrial robots in Norway, it is also useful to get some descriptive figures from Eurostat' survey on "ICT usage and e-commerce in enterprises (2018)" (see Tables A1 to A4 in S1 File). Large firms are the main adopters of both industrial and service robot technologies, and capital-intensive firms appear to invest in and integrate both technologies in their operations. Operating machines represent about 60% of all industrial robots in Norwegian firms in 2017. Whereas large firms use service robots for mostly logistics and transportation purposes, small and medium enterprises (SMEs) deploy robots in more product-related purposes, such as inspection, assembly or construction works.

Although our paper focuses on industrial automation, workers in knowledge-intensive service occupations may rather fear competition from new artificial intelligence technologies. Table A4 in S1 File presents some descriptives on Norwegian firms' use of Big Data in their business operations. Large firms are more likely to use Big Data than SMEs. Large firms use smart sensors (e.g. Internet of Things) and geo-data to a greater extent than SMEs. On the other hand, SMEs more actively collect data from social media for marketing purposes. In sum, smart machines are swiftly making inroads in the Norwegian economy, and this pace has accelerated in the last five years.

Fig 3 shows the time trend of the variable machine replacement for each of the 16 industry-regions in the more recent period 2016–2019 to which our survey data refers. Although this is a relatively short span (which does not make it possible to assess long run trends), Fig 3 indicates that fear of replacement due to automation has increased steadily in most of the industry-regions considered in this study, and that there is by and large a correspondence between the time trends reported in Fig 2 (robot adoption) and Fig 3 (machine replacement).

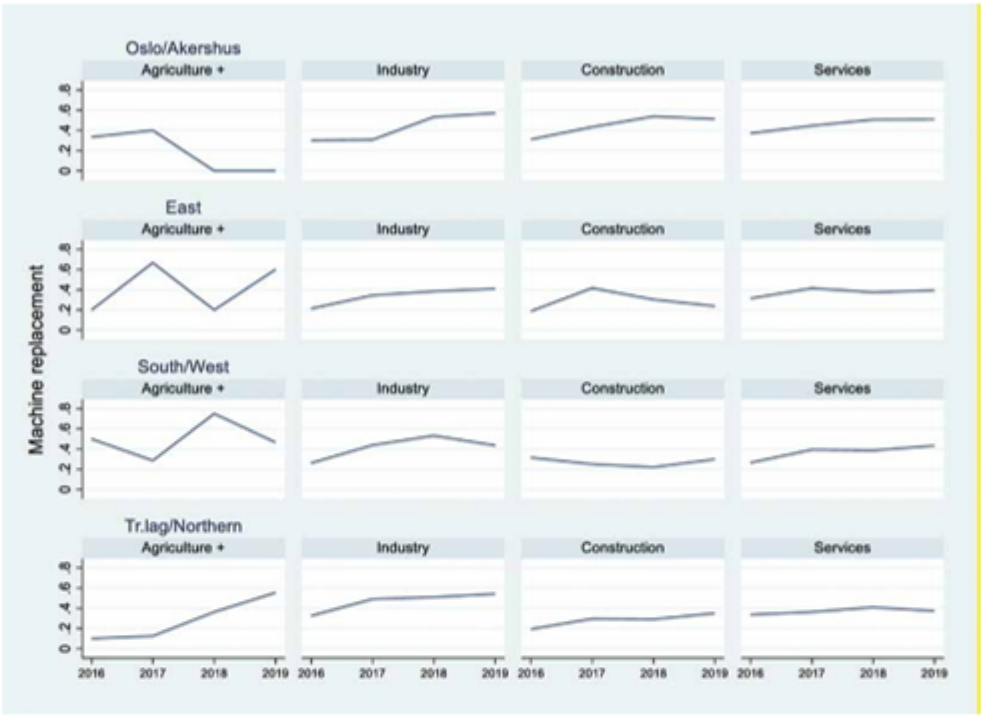


Fig 2.6 Share of workers who believe that their job can be replaced by machines, by region and industry.
https://doi.org/10.1371/journal.pone.0242929.g003

Based on the identification strategy noted above, we estimate a two-stage instrumental variables (IV) model: the first stage (3) investigates how robot exposure and other control factors explain variations in workers' fear of replacement, whereas the second stage (2) estimates the relationship between job satisfaction and anticipated replacement:

$$JS_{irst} = \alpha_1 + \gamma machine\ replacement_{irst} + \delta x'_{irst} + \eta_r + \theta_t + \varepsilon_{irst} \quad (2)$$

$$machine\ replacement_{irst} = \alpha_2 + \mu z_{rst} + \rho x'_{irst} + \tau_r + \varphi_t + \in_{irst} \quad (3)$$

JS is reported job satisfaction, *machine replacement* is the dummy variable indicating whether the respondent believes a machine can perform her/his job tasks, *z* is the instrumental variable (industry-region lagged pace of robot adoption), and *x* is a set of covariates (measured for individuals in each survey wave). The subscript *r* denotes the geographical region of residence of each worker *i*, *s* denotes the industry in which the worker is employed, and the subscript *t* refers to survey year. Among the set of covariates, the skill variable is particularly relevant for the present study, as we seek to investigate whether the relationship between fear of replacement and job satisfaction differs for high- *versus* low-skilled workers. To test these moderation effects, we interact the skill variable with the robots variable in the first stage equation, and with the fear of replacement variable in the second stage equation.

For model identification, the vector x in Eqs (2) and (3) does also include detailed demographic and socio-economic characteristics expected to correlate with job satisfaction and anticipated replacement, such as age, gender, income, union membership, and industry employment. According to previous studies, these factors are relevant to explain variation in job satisfaction, labor dynamics and technological automation diffusion [8,22,31,34,37,43–45]. Finally, both equations also include a full set of regional dummies and time dummies that control for unobservable determinants of job satisfaction within each region over time.

It is important to note that our identification strategy is based on the assumption that robot exposure in each industry-region will affect workers' job satisfaction only through its effects on fear of replacement, and we therefore exclude a direct impact of robot exposure on job satisfaction. Conceptually, we cannot exclude that robot adoption in a given firm may potentially affect employees' job satisfaction directly, and not only indirectly through fear of replacement. However, we think that this conceptual argument is not a particular reason of concern in our empirical study. The reason for this is that robot adoption in Norway, although it has increased rapidly during the last few years, it is still relatively low in absolute levels (around 6%, see Table 2). This means that our dataset and estimations do not refer to workers who already use robots in their current job, but for the great majority they refer to workers that are exposed to (i.e. observe) automation being introduced in other firms in the industry-region where they work, and that due to these peer effects fear that machines could replace some of their working tasks in the future.

The econometric model is estimated as a two-stage bivariate recursive ordered probit maximum likelihood setup, which accommodates the ordinal character of the outcome and main explanatory variable [46,47]. This model estimates response probabilities of two variables, one ordered and one dichotomous, and the exogenous variable robot exposure is included in the first stage [48,49]. Estimations are performed with Roodman's [50] conditional mixed process (CMP) program. Because the instrument is measured at the industry-region level, estimations are likely to contain grouped structures, and we therefore cluster standard errors in all regressions [41,51,52].

6. Results

First Stage Results

Table 4 presents the estimation results of the first stage (Eq 3), in which the dependent variable is machine replacement (i.e. workers' self-reported assessment of the possibility that some of their working tasks will be replaced by a smart machine in the future). Table 4 reports estimation results for both the model without control variables and the one including the full set of controls, in order to assess whether the inclusion of controls affect the results [53,54]. The results for the two models are however very close to each other. We begin by briefly looking at the results for the set of control variables, before turning attention to the main variables of our interest. Among the controls, Table 4 shows that fear of replacement is stronger for younger workers. These have a longer time horizon remaining in their working life, and they are therefore more likely to expect that automation will replace some of their working tasks in the future [12]. Employees that belong to a trade union are less likely to fear replacement, arguably because their employment and working conditions are partly protected by the trade union membership (we elaborate further on this in section 6 below). Regarding wage levels, fear of replacement is stronger for workers that have higher income. A possible interpretation of this finding is that, after controlling for education and skill levels (that are correlated with wages and that also affect fear of replacement), workers with higher income have more to lose *vis-à-vis* workers with lower income, since automation of tasks may lead, in absolute terms, to a stronger wage decrease for them. Further, we control for gender and sector of occupation (industry employment), which are two standard control variables in studies of workers' subjective well-being.

Table 2.8 First stage results. Dependent variable: Machine replacement.

	(1)	(2)	(3)	(4)
Robot adoption	0.711***	0.853***	1.845***	2.250***
	(0.217)	(0.256)	(0.570)	(0.681)
Age		-0.004***		-0.011***
		(0.001)		(0.002)
Union membership		-0.056***		-0.149***
		(0.013)		(0.035)
Income scale = 2		-0.093***		-0.268***
		(0.024)		(0.065)
Income scale = 3		-0.006		-0.019
		(0.022)		(0.057)
Income scale = 4		0.024		0.063
		(0.023)		(0.060)
Income scale = 5		0.027		0.072
		(0.029)		(0.077)
Income scale = 6		0.043		0.112
		(0.034)		(0.090)
Income scale = 7		0.078***		0.206***
		(0.026)		(0.069)
Income scale = 8		0.039		0.100
		(0.026)		(0.070)
Income scale = 9		0.069		0.182
		(0.053)		(0.137)
University degree		0.037*		0.098*
		(0.021)		(0.055)
Woman		-0.011		-0.030
		(0.017)		(0.044)
Industry employment		-0.044**		-0.119***
		(0.016)		(0.042)
Controls				
Regional dummies	✓	✓	✓	✓
Year dummies	✓	✓	✓	✓
F-stat	10.77	11.15		
N	10,051	10,051	10,051	10,051

Robust standard errors in parentheses are clustered for workers in the same region and industry. Columns 1 and 2 present OLS estimates. Columns 3 and 4 show probit estimates.

* p<0.10,

** p<0.05,

***p<0.01.

https://doi.org/10.1371/journal.pone.0242929.t004

Shifting the focus to the main variables of interest for this study, the instrumental variable (changes in robot adoption in local labor markets between 2008 and year *t*) is as expected positively correlated with the dependent variable (workers' fear of replacement). As explained in section 5, the underlying idea is that when individuals work in industry-regions in which firms have increasingly been using robots in the last few years, they are more *exposed* to automation (e.g. because some of their peers or acquaintances in the same region have lost their job due to automation). These peer effects translate into fear of replacement even for workers that are still employed and not directly touched by automation technologies yet. Table A6 in S1 File corroborates this interpretation by presenting first stage regressions in which we use two additional control variables that measure *fear of job loss*: (1) job loss worry ("To what extent are you worried about losing your job?"); (2) unemployed in five years ("Do you expect to be unemployed within the next five years?"). Both of these control variables are positive and significant in the regressions, indicating that fear of job loss (emotional reaction) and machine replacement (cognitive reaction) are closely related dimensions. Further, the inclusion of these additional control variables does not affect the size and significance of the estimated effect of the instrumental variable robot adoption on machine replacement.

What is the economic significance of these results? According to our OLS estimates (see column 2 in Table 4), a one standard deviation increase in robot exposure increases the probability that a worker expresses fear of machine replacement by 2.6 percent (we thank an anonymous reviewer for suggesting to point this out). It is hard to say whether this estimated effect is economically sizeable. However, considering that robot adoption in Norway has more than doubled during the time span considered in this study, we may think that the overall effect of automation on fear of replacement has arguably been important during this period. It is also interesting to assess this figure in the light of the effects of robot adoption on employment found in other recent studies (although none of these present estimates of the effects of robot adoption on subjective fear of replacement). [34] find that one additional robot per thousand workers reduces the employment rate by 0.16–0.20 percentage points across European regions. In their study of the U.S. labor market, [22] find that one robot reduces employment by three workers. Further, [36] suggest that the corresponding number for German manufacturing workers is about two jobs. However, the overall effect on German employment is unchanged as the job reduction in manufacturing is offset by gains in the service sector.

Next, we look at the results for the other important variable considered in this study: workers' skills. Table 4 shows that individuals with tertiary education have on average a greater fear that some of their working tasks could be replaced by a machine in the future. As discussed in section 3, this might be explained by the fact that high-skilled workers are typically more educated individuals who read more and follow societal debates on the media about robots, AI and automation, and their negative consequences for employment. Hence, high skilled workers are arguably more exposed to peer effects, which may translate in a greater fear about the future of employment.

Relatedly, how do workers' skills affect the positive relationship between automation and fear of replacement? To test this moderation effect, Table 5 reports estimation results of the first stage equation by workers' skill level. While the estimated coefficient of the robot adoption variable is positive and significant for both workers with tertiary education and those without a college degree, the size of this effect is larger for the latter group. This moderation effect is in line with the recent literature on the effects of automation on employment, which shows that low-skilled workers are more exposed to the risks of displacement from automation because they typically carry out routine tasks that can more easily be automated [3,4,16,20].

Table 2.9 First stage results by workers' skill level.

	(1)	(2)
	No university education	**University education**
Robot adoption	1.040***	0.749**
	(0.306)	(0.283)
Age	-0.004***	-0.004***
	(0.001)	(0.001)
Union membership	-0.024	-0.085***
	(0.020)	(0.020)
Income scale = 2	-0.111**	-0.048*
	(0.042)	(0.023)
Income scale = 3	0.005	-0.024
	(0.052)	(0.036)
Income scale = 4	0.037	0.031
	(0.050)	(0.018)
Income scale = 5	0.092	0.010
	(0.070)	(0.024)
Income scale = 6	0.024	0.067*
	(0.047)	(0.034)
Income scale = 7	0.099	0.080**
	(0.066)	(0.033)
Income scale = 8	0.061	0.032
	(0.051)	(0.035)
Income scale = 9	0.072	0.068
	(0.056)	(0.062)
Woman	0.052**	-0.058***
	(0.023)	(0.018)
Industry employment	-0.057***	-0.005
	(0.018)	(0.030)
Controls		
Regional dummies	✓	✓
Year dummies	✓	✓
N	4,434	5,617

Robust standard errors in parentheses are clustered for workers in the same region and industry. Columns 1 and 2 present OLS estimates.

* p<0.10,

** p<0.05,

*** p<0.01.

https://doi.org/10.1371/journal.pone.0242929.t005

Second Stage Results

Table 6 reports estimation results for the second stage of the model (Eq 2), in which job satisfaction is the dependent variable. The table reports first the results for the model without control variables and then those for the model including the full set of controls. The results for the two models are very close to each other, indicating that the inclusion of controls does not affect the results on the main explanatory variables [53,54]. The control variables that we use are commonly used in the job satisfaction literature. Income is positively correlated with job satisfaction, in line with extant literature showing that wage is one of the factors that enhance workers' subjective well-being [7,8]. Female workers report on average higher job satisfaction than males; and individuals employed in manufacturing (industry) have lower satisfaction than average, a finding that is explained in the subjective well-being literature by the fact that factory workers typically carry out routine and monotonous working tasks and have a lower degree of autonomy and creativity [25].

Table 2.10 Second stage results. Dependent variable: Job satisfaction.

	(1)	(2)	(3)	(4)
Machine replacement	-1.268**	-0.760**	-1.093***	-0.999***
	(0.512)	(0.304)	(0.117)	(0.169)
Age		0.007***		0.008***
		(0.001)		(0.002)
Union membership		-0.041*		-0.057***
		(0.024)		(0.020)
Income scale = 2		-0.135**		-0.137***
		(0.067)		(0.052)
Income scale = 3		0.009		0.016
		(0.037)		(0.038)
Income scale = 4		0.093*		0.120**
		(0.053)		(0.054)
Income scale = 5		0.144***		0.187***
		(0.044)		(0.049)
Income scale = 6		0.195***		0.246***
		(0.057)		(0.066)
Income scale = 7		0.214***		0.285***
		(0.049)		(0.053)
Income scale = 8		0.203***		0.274***
		(0.060)		(0.073)
Income scale = 9		0.322***		0.434***
		(0.071)		(0.094)
University degree		0.035		0.034
		(0.024)		(0.028)
Woman		0.092***		0.129***
		(0.020)		(0.028)
Industry employment		-0.169***		-0.213***

(*Continued*)

https://doi.org/10.1371/journal.pone.0242929

Table 2.10 Second stage results. Dependent variable: Job satisfaction. ***(Continued)***

	(1)	(2)	(3)	(4)
		(0.020)		(0.026)
Controls				
Regional dummies	✓	✓	✓	✓
Year dummies	✓	✓	✓	✓
N	10,051	10,051	10,051	10,051

Robust standard errors in parentheses are clustered for workers in the same region and industry. Columns 1 and 2 present 2SLS linear estimates. Columns 3 and 4 show bivariate recursive probit estimates.
* p<0.10,
** p<0.05,
***p<0.01.
https://doi.org/10.1371/journal.pone.0242929.t006

The main variable of interest in Table 6 is machine replacement. The estimated coefficient for this variable is as expected negative and significant. This means that workers that report higher fear of replacement from smart machines have on average lower job satisfaction. As noted in section 3, the reason for this is that for workers that are currently employed, the prospect that smart machines could replace some of their current working tasks in the future, or even the whole job, does create uncertainty about future job prospects and personal finance, thus lowering job satisfaction. The marginal effects for each category of the job satisfaction variable (not reported in Table 6) indicate that a change in the fear of replacement dummy variable (i.e. from "no fear" to "fear of replacement") decreases the job satisfaction of the "very satisfied" workers by around 30%, and of the "very or pretty dissatisfied" workers by around 6–11%.

Table A7 in S1 File reports second stage regressions that also include two additional control variables that measure *fear of job loss*: (job loss worry; unemployed in five years; see definition of these two survey questions in section 6 above). The additional control variables are positive and significant in the regressions, and their inclusion in the model does not affect the size and significance of the estimated effect of machine replacement on job satisfaction, indicating that fear of replacement due to automation is important for workers' subjective well-being even after controlling for the more general construct fear of job loss.

How is this relationship moderated by workers' skill level? Table 7 investigates this question by reporting marginal effects of the machine replacement variable for workers that have tertiary education *versus* those that do not have a college degree. The table shows that the marginal effect for the workers without a college degree is negative and significant, indicating that fear of replacement increases the probability that low-educated workers will report high job satisfaction by nearly 50%. On the other hand, the corresponding marginal effect for the workers with a University degree is not statistically significant. Abeliansky and Beulmann [11] also carried out some regressions to study the relationships between automation and mental health for different educational groups (tertiary vs secondary education), but they did not find any significant difference among these groups of workers in Germany. As discussed in section 3, the interpretation of our finding is that low-skilled workers, if replaced, will on average have fewer possibilities to find another occupation in the labor market. This is in line with recent literature that provides evidence that displacement effects of industrial robots on employment and wages are stronger and more significant for low-education workers [20,22]. On the other hand, automation technologies can have more positive effects on high-skilled workers, increasing the demand for labor, wages and the complexity and interest of their tasks [18]. In short, we posit that workers are at least to some extent aware of the distinct impacts that automation can have for different types of occupations, and this explains why fear of replacement turns out to be a concern for low-skilled employees.

Table 2.11 Second stage results by workers' skill level (marginal effects of machine replacement for workers of different education levels).

	Below university education	University education
Machine replacement	-0.494***	0.295
	(0.165)	(0.683)
Controls		
Individual controls	✓	✓
Regional dummies	✓	✓
Year dummies	✓	✓
N	10,051	10,051

Robust standard errors in parentheses are clustered for workers in the same region and industry. Columns 1 and 2 present results from bivariate recursive probit estimates.
* p<0.10,
** p<0.05,
*** p<0.01.
https://doi.org/10.1371/journal.pone.0242929.t007

It may be argued that the education level dummy variables that we have used in these regressions only reflect formal education acquired through the school and University system, and disregards other skills that workers acquire during the working life through training, apprenticeships and learning by doing. Ideally, if we had information about each worker's occupation, we could construct a proxy measure for skills by using the three-level job complexity schema developed by Hunter et. al. [55], which creates a correspondence between job types and corresponding skill content. However, our dataset does not have information about workers' occupation type, and we are therefore not able to follow this route. Hence, in the absence of a more specific variable measuring workers' skills, we carry out two additional exercises. First, we use age as an additional proxy of workers' skills and abilities to perform their job. Table 8 reports marginal effects of machine replacement on job satisfaction for workers of different education levels *and* for different age groups. The results confirm the main finding noted above. The marginal effect of anticipated machine replacement on present job satisfaction of the workers without a college degree is negative and significant for all sub-groups (except those younger than 30), and it is not significant for workers with a University degree. This corroborates the main finding of our analysis that machine replacement has a negative effect on job satisfaction, and that this effect is particularly relevant for low-skilled workers.

Table 2.12 Second stage results by workers' skill level and age (marginal effects of machine replacement for workers of different education levels and different age groups).

	Below university education	University education
Age group		
<30 years	0.109	1.552
	(0.363)	(2.205)
30–44 years	-0.427**	-1.306
	(0.171)	(0.871)
45–59 years	-0.385*	0.138
	(0.205)	(0.443)
60+ years	-1.180***	-0.486

(*Continued*)

Table 2.12 Second stage results by workers' skill level and age (marginal effects of machine replacement for workers of different education levels and different age groups). (*Continued*)

	Below university education	University education
	(0.235)	(1.478)
Controls		
Individual controls	✓	✓
Regional dummies	✓	✓
Year dummies	✓	✓
N	10,051	10,051

Robust standard errors in parentheses are clustered for workers in the same region and industry. Columns 1 and 2 present results from bivariate recursive probit estimates.
* p<0.10,
** p<0.05,
***p<0.01.
https://doi.org/10.1371/journal.pone.0242929.t008

Second, it may be argued that the education variable does not only measure workers' skills, but it is also a proxy for *employability*, since workers with higher education levels can more easily find a new job. If so, employability, rather than skills, could be the latent variable moderating the effect of fear of replacement on job satisfaction. To address this possibility, we make use of two additional variables measuring employability: (1) Difficult to find a new job ("How difficult or easy do you think it would be for you to find a job that is at least as good as the one you have now?"); (2) Insufficient skills in current job ("How often do you experience insufficient competence to perform your tasks?"). Then we include these two variables as additional controls in our first and second stage regressions, and report the results of these robustness tests in tables A8 to A10 in S1 File. First stage results (Tables A8 and A9 in S1 File) show that the inclusion of the additional controls for employability does not affect the main result about the effect of robot adoption on machine replacement, and that this effect is still stronger for workers with lower education level. Second stage results (Table A10 in S1 File) are also in line with our baseline estimations: in the extended model specification that controls for employability, the effect of machine replacement on job satisfaction is still negative and significantfor workers of lower education level (across age groups). In short, these additional exercises show that, even when we control for employability, workers' education level moderates the effect of fear of replacement on job satisfaction, and it may thus be considered as a reasonable proxy measure of latent workers' skills.

Robustness Tests

Our identification strategy rests on the assumption that the (lagged) introduction of robots in local labor markets in Norway affects current job satisfaction only through its effects on workers' fear of replacement. Although our regressions control for a set of relevant employee-level characteristics and include region- and time fixed effects, it is also useful to carry out some additional robustness exercises to test the potential confounding effects of omitted variables that may in principle affect both fear of replacement and the error term of the outcome equation.

Tables A11 and A12 in S1 File report estimation results of first and second stage regressions that include some additional region-level control variables in the model. The first two columns add region's GDP and tertiary education level, which may be thought to be general relevant factors that may drive both the introduction of industrial robots and job satisfaction patterns. Though, the estimated coefficients of the instrumental variable robot adoption (Table A11 in S1 File) and of the machine replacement variable (Table A12 in S1 File) are still significant and stable after the introduction of these two possible confounding factors. Regressions in

column 3 add a variable measuring business building broadband infrastructure in each region. The reason for including this variable is that ICT diffusion may be a potentially conflating factor that can disturb the effect of robots adoption on employment [3,37]. By controlling for broadband internet access in office buildings we address this concern, reasonably assuming that the development of broadband infrastructure is driven by policies and investments that are exogenous to the individual worker. Again, the inclusion of this additional control does not affect the estimated coefficient of the robot variable in Table A11 in S1 File, and of the machine replacement variable in Table A12 in S1 File. These coefficients still have the same signs and significance levels, and their estimated size is slightly larger than in baseline regressions. Finally, columns 4, 5 and 6 also add three other region-level controls: unemployment benefit recipients (share of population in each region), share of large companies in each region, and population size (log). The unemployment benefit variable controls for the possible confounding effect of different unemployment rates across regions. The share of large companies takes into account the fact that large firms do on average have a higher rate of introduction and use of industrial robots (see Tables A1 and A4 in S1 File), so that employees in regions with a high share of large firms are potentially more exposed to the effects of automation. Finally, the population variable is a standard control for the size and density of the region, which may be related to the extent and intensity of peer effects that affect workers' fear of replacement. However, the inclusion of these additional variables does not affect the main results for the explanatory variables of our interest.

As a further robustness test, Table A13 in S1 File reports the results of a placebo test that adds a lead variable–robust exposure at year $t+1$ –to the set of regressors in the first stage equation (including also the three additional region-level control variables noted in the previous paragraph). In these placebo regressions, the future robot adoption variable is not significant, and its inclusion does not affect the sign and size of the estimated coefficient of the instrumental variable (lagged pace of robot adoption). This further rules out the possibility that our results are driven by some omitted variables that are related to both job satisfaction and robot adoption.

Next, it is interesting to consider the question on whether our instrumental variable (pace of robot adoption in industry-regions) should be regarded as a *peer effect* or rather a *neighbor effect*. Table A14 (in S1 File) considers this question by including two additional control variables in the first stage regressions. The first additional control is robot adoption in the other three industries in the same region; and the second one is robot adoption in the same industry in all other regions in Norway. In other words, these two additional controls are proxies for inter-industry and inter-regional neighboring effects, respectively. Table A14 in S1 File shows that none of these variables are significantly correlated with fear of replacement, and their inclusion does not affect the precision of the instrumental variable (robot adoption in a given industry-region). These robustness exercises provide further evidence that our instrumental variable catches peer effects that are specific to each industry-region, and that hold even after controlling for neighbor effects related to robot adoption in surrounding regions and industries in Norway.

Finally, it is relevant to comment further on the role of one of the control variables in the model: union membership. As noted in relation to Table 4 (and other first stage results reported in the online appendix), workers that belong to a trade union do on average report lower fear of machine replacement. This may suggest that workers in trade unions feel they are more protected from the impacts of industrial robots. However, this pattern is in contrast with Acemoglu and Restrepo [44], which find a positive association between industrial robot adoption and unionization rates across countries, arguing that this is due to the fact that unionization may raise labor costs. Yet, skill-biased technical change also creates a stronger incentive for deunionization because the outside employment and wage options of skilled workers have improved [56]. To investigate this further, we run additional regressions in which we interact our two main explanatory variables (robot adoption in the first stage, and machine replacement in the second stage) with the union membership variable. The idea is to test directly whether Norwegian workers that belong to a trade union do on average

think that they are less likely to be affected by automation. However, the results of these regressions (reported in Table A14 in S1 File) show that the two additional interaction variables are not statistically significant. We think that the role of union membership as a factor moderating the effects of industrial automation is an interesting topic for future research.

7. Conclusions

The swift pace of introduction of industrial robots, AI and smart machines in production activities in recent years represents a new major process of Schumpeterian creative destruction. This process will in the near future lead to dramatic consequences for employment in many sectors and regions, and it will at the same time create new unprecedented opportunities for productivity growth, wealth and well-being. As for other major transformations in the past, this structural change and the related transition and adjustment process will arguably not be smooth and swift: it will unfold over a period of several years, and it will lead to important negative impacts in the short-run before the long-run economic and societal benefits will eventually emerge.

Studying the effects of automation on employment, extant research has so far mostly focused on aggregate impacts that industrial robots and AI have on employment demand and wages for different industries and countries. The present paper has argued that it is important to shift the focus to the micro-level of analysis and study the impacts of automation technologies on individual workers' well-being. Specifically, we have put forward the idea that the relevant impacts that it is important to study are not only pecuniary (i.e. related to workers' employment conditions and wages) but also nonpecuniary (i.e. related to workers' expectations and future job prospects). *Ceteris paribus*, workers that fear that their working tasks might be replaced by a smart machine in the future may have a lower job satisfaction at present than workers who have more secure job prospects and less uncertainty about the future.

We have investigated this idea by considering a large sample of workers in Norway for the period 2016–2019, and studying the extent to which the introduction of industrial robots in local labor markets affect workers' fear of being replaced in the future, and in this way hamper their subjective well-being. Our data and results provide a quite striking picture. 40% of Norwegian workers in our sample think their working tasks might be replaced by a machine, and our analysis shows that this fear of replacement significantly lowers their job satisfaction at present. We also find that this transmission mechanism is driven by low-skilled workers, which are those carrying out routine-based tasks, and who are therefore aware to be more exposed to the risks of automation. On the whole, we think that our empirical findings are not only relevant for Norway (the country to which our dataset refers), but they can in principle have more general lessons for other countries too. Automation is by now an important trend that is rapidly diffusing worldwide, and its effect on workers' health and well-being is therefore a topic of high societal relevance. Schwabe (2019) provides related evidence using a different dataset for a larger sample of European countries. The present work calls therefore for further research that may investigate and extend this research topic in a variety of different countries and continents.

A first important policy implication of our results is that the current process of structural change and creative destruction will in the short-run likely lead to stronger fear of replacement and uncertainty about the future for low-skilled workers carrying out routine work in factories, thus possibly leading to further polarization not only in terms of employment and wages, but also in terms of subjective well-being. To mitigate these negative consequences, which are already visible at present, national authorities should actively support training and re-training policies in such a way that workers that are exposed to future replacement may build up new competencies that can increase their ability to work with smart machines, as well as increase their qualifications and the likelihood to find a new job if this will become necessary in the future. If fear of replacement triggers workers to participate in such training is an interesting question for future studies. In other words, by giving better future prospects to more vulnerable workers, training policies will also contribute to enhance their subjective well-being at present.

https://doi.org/10.1371/journal.pone.0242929

Our results also suggest a second reflection and possible policy implication. As noted above, 40% of Norwegian workers in our sample think that their working tasks might be replaced by a machine. According to the Eurobarometer survey, the extent of fear of replacement is roughly the same for workers in other European countries [12]. This number is quite high indeed. Is it reasonable that so many workers fear competition from smart machines, and why is it so?

Extant research on automation and employment has not yet reached a consensus on the direction and size of these effects, and it still presents a vivid debate between those that emphasize negative consequences and those that point out positive economic and societal effects. Hence, there is no clear scientific evidence and consensus at present that could provide the basis for individual workers to form rational and well-informed assessments and expectations about their job prospects in the future. It is therefore reasonable to ask whether the generalized fear of competition from smart machines is actually exaggerated and not based on extant research and established knowledge. The concrete risk is that–in the current phase of rapid and disruptive technological change–societal debates in the media on robots, automation and AI may tend to exaggerate risks and depict gloomy future scenarios, while often neglecting possible long-run benefits for the economy and the society, which are indeed even hard to imagine at the moment [1].

Since media debates on this topic are often biased and tend to overemphasize the negative impacts of automation (which are arguably more "catchy" and attractive for the uninformed audience), this may contribute to explain why so many workers report to fear future machine replacement. However, our paper has shown that such subjective individual assessments about the future may indeed hamper job satisfaction at present. This can also lead to anxiety, mental stress and low motivation at work, which may in turn depress creativity, productivity and innovation in the workplace.

In short, we should not disregard the possibility that a biased and uninformed presentation of this topic in the media may indeed have concrete negative consequences on workers' subjective well-being by affecting their beliefs about future job prospects. The policy implications of this are certainly not easy to draw. A major point, though, is to stress the importance of having better informed societal debates in the media, and particularly in State-owned channels, that take a more balanced view of the negative and positive consequences of automation, and that avoid spreading fears and gloomy scenarios that are not based on solid evidence and arguments.

Supporting Information

S1 File. Online appendix: Additional information and robustness tests. Containing Tables A.1 to A.15.
(DOCX)

Author Contributions

Conceptualization: Henrik Schwabe, Fulvio Castellacci.
Data curation: Henrik Schwabe.
Formal analysis: Henrik Schwabe.
Funding acquisition: Fulvio Castellacci.
Investigation: Fulvio Castellacci.
Methodology: Henrik Schwabe, Fulvio Castellacci.

References

1. Mokyr J., Vickers C. and Ziebarth N.L., The History of Technological Anxiety and the Future of Economic Growth: Is This Time Different? Journal of Economic Perspectives, 2015. 29(3): p. 31–50.
2. Frank M.R., et al., Toward understanding the impact of artificial intelligence on labor. Proceedings of the National Academy of Sciences, 2019. 116(14): p. 6531–6539. https://doi.org/10.1073/pnas.1900949116 PMID:30910965
3. Autor D.H. & Dorn D., The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market. The American Economic Review, 2013. 103(5): p. 1553–1597.
4. Goos M., Manning A. and Salomons A., Explaining Job Polarization: Routine-Biased Technological Change and Offshoring. American Economic Review, 2014. 104(8): p. 2509–26.
5. Acemoglu, D. & P. Restrepo, Artificial Intelligence, Automation and Work. NBER Working Paper (No. w24196), 2018.
6. Kaplan G. & Schulhoffer-Wohl S., The Changing (Dis-)Utility of Work. Journal of Economic Perspectives, 2018. 32(3): p. 239–58. PMID:30362697
7. Erdogan B., et al., Whistle While You Work: A Review of the Life Satisfaction Literature. Journal of management, 2012. 38(4): p. 1038–1083.
8. Clark A., Oswald A. and Warr P., Is job satisfaction U-shaped in age? Journal of occupational and organizational psychology, 1996. 69(1): p. 57–81.
9. Böckerman P., Ilmakunnas P. and Johansson E., Job security and employee well-being: Evidence from matched survey and register data. Labour Economics, 2011. 18(4): p. 547–554.
10. Oswald A.J., Proto E. and Sgroi D., Happiness and Productivity. Journal of Labor Economics, 2015. 33(4): p. 789–822.
11. Abeliansky, A.L. & M. Beulmann. (2019). Are they coming for us? Industrial robots and the mental health of workers. cege Discussion Papers, No. 379. University of Göttingen, Center for European, Governance and Economic Development Research (cege), Göttingen.
12. Schwabe, H., Automation, Fear of Replacement and the Subjective Well-Being of Workers. TIK working paper, 2019.
13. Rajan K. & Saffiotti A., Towards a science of integrated AI and Robotics. Artificial Intelligence: Special Issue on AI and Robotics, 2017. **247**: p. 1–9.
14. Acemoglu, D. & D. Autor, Skills, tasks and technologies: Implications for employment and earnings, in Handbook of labor economics. 2011, Elsevier. p. 1043–1171.
15. Sachs, J.D. & L.J. Kotlikoff. (2012). Smart Machines and Long-Term Misery. NBER Working Paper (No. w18629). National Bureau of Economic Research.
16. Autor D.H., Why Are There Still So Many Jobs? The History and Future of Workplace Automation. Journal of Economic Perspectives, 2015. 29(3): p. 3–30.
17. Houseman S.N. (2018). Understanding the decline of US manufacturing employment. W.E. Upjohn Institute for Employment Research. Upjohn Institute working paper.
18. McGuinness, S., K. Pouliakas and P. Redmond, Skills-Displacing Technological Change and Its Impact on Jobs: Challenging Technological Alarmism? IZA Discussion Paper No. 12541, 2019.

19. Klenert, D., E. Fernandez-Macias and J.-I. Anton. (2020). Do robots really destroy jobs? Evidence from Europe. Seville: European Commission. Joint Research Centre.

20. Blanas S., Gancia G. and Lee S.Y.T.. (2019). Who is afraid of machines? Barcelone GSE Working Paper Series.

21. Beaudry P., Green D.A. and Sand B.M., The great reversal in the demand for skill and cognitive tasks. Journal of Labor Economics, 2016. 34(S1): p. S199–S247.

22. Acemoglu D. & Restrepo P., Robots and Jobs: Evidence from US Labor Markets. Journal of Political Economy, 2020. 128(6): p. 2188–2244.

23. Acemoglu, D. & P. Restrepo. (2020). Unpacking Skill Bias: Automation and New Tasks. National Bureau of Economic Research.

24. Bessen, J., et al. (2020). Automation: A Guide for Policymakers. Retrieved from https://www.brookings.edu/wp-content/uploads/2020/01/Bessen-et-al_Full-report.pdf.

25. Castellacci F. & Viñas-Bardolet C., Internet use and job satisfaction. Computers in Human Behavior, 2019. 90: p. 141–152.

26. Green F. & Zhu Y., Overqualification, job dissatisfaction and increasing dispersion in the returns to graduate education. Oxford economic papers, 2010. 62(4): p. 740–763.

27. Sousa-Poza A. & Sousa-Poza A.A., Well-being at work: a cross-national analysis of the levels and determinants of job satisfaction. The journal of socio-economics, 2000. 29(6): p. 517–538.

28. Salvatori A., Labour contract regulations and workers' wellbeing: International longitudinal evidence. Labour Economics, 2010. 17(4): p. 667–678.

29. Kahneman D. & Tversky A., Prospect Theory: An Analysis of Decision under Risk. Econometrica, 1979. 47(2): p. 263–291.

30. Belfield C.R. & Harris R.D., How well do theories of job matching explain variations in job satisfaction across education levels? Evidence for UK graduates. Applied economics, 2002. 34(5): p. 535–548.

31. Aghion P., et al., Creative destruction and subjective well-being. American Economic Review, 2016. 106(12): p. 3869–97. https://doi.org/10.1257/aer.20150338 PMID:28713168

32. IFR. (2017). World Robotics: Industrial Robots. International Federation of Robotics.

33. IFR. (2018). World Robotics 2018: Industrial Robots. Retrieved from International Federation of Robotics: https://ifr.org/.

34. Chiacchio, F., G. Petropoulos and D. Pichler, The impact of industrial robots on EU employment and wages: A local labour market approach. Bruegel Working Papers (02), 2018.

35. Graetz G. & Michaels G., Robots at work. The Review of Economics and Statistics, 2018. 100(5): p. 753–768.

36. Dauth, W., et al., Adjusting to Robots: Worker-Level Evidence. Opportunity and Inclusive Growth Institute Working Paper (13), 2018.

37. Michaels G., Natraj A. and Van Reenen J., Has ICT Polarized Skill Demand? Evidence from Eleven Countries over Twenty-Five Years. Review of Economics and Statistics, 2014. 96(1): p. 60–77.

38. Falk M. & Biagi F., Relative demand for highly skilled workers and use of different ICT technologies. Applied Economics, 2017. 49(9): p. 903–914.

39. Bhuller M., et al., Broadband internet: An information superhighway to sex crime? Review of Economic Studies, 2013. 80(4): p. 1237–1266.

40. Akerman A., Gaarder I. and Mogstad M., The skill complementarity of broadband internet. The Quarterly Journal of Economics, 2015. 130(4): p. 1781–1824.

41. Angrist J.D. & Pischke J.-S., Mostly harmless econometrics: An empiricist's companion. Vol. 2009, Princeton, NJ.: Princeton university press.

42. European Commission. Digital Single Market. 2014; Available from: https://europa.eu/european-union/file/1497/.

43. Frank M.R., et al., Small cities face greater impact from automation. Journal of the Royal Society Interface, 2018. 15(139): p. 20170946. https://doi.org/10.1098/rsif.2017.0946 PMID:29436514

44. Acemoglu, D. & P. Restrepo, Demographics and Automation. NBER Working Paper (No. w24421), 2018.

45. Card D., et al., Inequality at work: The effect of peer salaries on job satisfaction. American Economic Review, 2012. 102(6): p. 2981–3003.

46. Sajaia Z., Maximum likelihood estimation of a bivariate ordered probit model: implementation and Monte Carlo simulations. The Stata Journal, 2008. 4(2): p. 1–18.

47. Monfardini C. & Radice R., Testing exogeneity in the bivariate probit model: A Monte Carlo study. Oxford Bulletin of Economics and Statistics, 2008. 70(2): p. 271–282.

48. Wooldridge, J.M., Econometric analysis of cross section and panel data. 2010: MIT press.

49. Maddala, G.S. & L.-F. Lee, Recursive models with qualitative endogenous variables, in Annals of Economic and Social Measurement, Volume 5, number 4. 1976, NBER. p. 525–545.

50. Roodman D., Fitting fully observed recursive mixed-process models with cmp. Stata Journal, 2011. 11(2): p. 159–206.

51. Moulton B.R., Random group effects and the precision of regression estimates. Journal of Econometrics, 1986. 32(3): p. 385–397.

52. Cameron C.A. & Miller D.L., A practitioner's guide to cluster-robust inference. Journal of Human Resources, 2015. 50(2): p. 317–372.

53. Becker T.E., Potential Problems in the Statistical Control of Variables in Organizational Research: A Qualitative Analysis With Recommendations. Organizational Research Methods, 2005. 8(3): p. 274–289.

54. Spector P.E. & Brannick M.T., Methodological Urban Legends: The Misuse of Statistical Control Variables. Organizational Research Methods, 2011. 14(2): p. 287–305.

55. Hunter J.E., Schmidt F.L. and Judiesch M.K., Individual differences in output variability as a function of job complexity. Journal of Applied Psychology, 1990. 75(1): p. 28–42.

56. Acemoglu, D., P. Aghion and G.L. Violante. Deunionization, technical change and inequality. in Carnegie-Rochester conference series on public policy. 2001. Elsevier.

2.10 Discussion Topics

The subject of automation and computer systems' effect on jobs is broad and complex and has numerous implications for the labor force and economy. These are a few of the areas that can be used to explore these subjects in more detail:

Historical Evolution of Automation in the Workplace

- Industrial Revolution to Digital Age: From the first machines of the Industrial Revolution to the present-day computer age, how has automation developed? The development of the first assembly lines, mechanization and computer-controlled processes can be mentioned.
- Impact on Job Types: What changes have occurred in the nature of work through automation in the movement from agriculture to manufacturing to services? How did technology influence the emergence of new industries alongside the disappearance of the old ones?

The Role of Computers in Modern Automation

- Introduction of Computers: The mid-20th century revolution in automation can be attributed to the introduction of computers. From manual, repetitive tasks to computer controlled processes in manufacturing, logistics, and other industries, describe the change.
- Software and AI in Automation: Software and artificial intelligence (AI) have extended the possibilities of automation. Discuss how AI has allowed decision making processes which were once the exclusive domain of people.

Impact on Employment and Job Displacement

- Job Displacement and Creation: Although automation has resulted in job losses in some areas, it has also generated new employment opportunities. What changes have occurred in employment status between job displacement and job creation within different industries? Which jobs are most at risk and which are expected to grow?
- Skills and Education Requirements: The shift towards automation has resulted in a change in the qualifications needed by the modern employee. Technical skills, continuous learning and adaptability to new technology are becoming essential in today's workforce.

Economic Implications of Automation

- Productivity and Economic Growth: Automation has been a source of increased productivity and economic development. What are the effects of increased productivity from automation on the general health of the economy and the distribution of wealth?
- Income Inequality: The positive impacts of automation are not shared by all. Explain how automation has contributed to income inequality, with owners of capital such as automated systems on one end and workers on the other.

Ethical and Social Considerations

- The Human-Machine Relationship: What are the ethical implications of increasingly sophisticated automation systems? Examine the moral issues of using machines to replace human work, especially in areas such as care, customer service and creative fields.
- Universal Basic Income (UBI): Some people propose UBI as a solution to the problem of job displacement due to automation. What is the effectiveness and potential effects of UBI in a world where automation is on the rise?

The Future of Work

- The Gig Economy and Freelancing: Traditional full-time jobs are being substituted by gig work and freelance work in the wake of automation. What are the consequences of this change for employment security, employee protection, and financial reliability?
- Human-Computer Collaboration: Instead of taking away humans, automation can enhance human potential. How can humans and machines work together in the workplace where AI and robots help with complicated work, while humans offer ideas, feelings and supervision?

Case Studies and Industry Examples

- Manufacturing and Robotics: Analyze how the automotive manufacturing industry has been transformed by robotics and automation. Discuss the changes in the labor force, production efficiency and competitiveness in the global market.
- Healthcare and Automation: Discuss how automation and computer systems are affecting the healthcare industry from diagnostic tools to robotic surgeries. What are the advantages and disadvantages of automation in a field that is so human?

Policy and Regulatory Challenges

- Regulation of Automation: How can governments control the deployment of automation to ensure that innovation does not harm society? What is the function of policy in steering the transition to an automated economy?
- Worker Retraining Programs need to establish programs which provide workers with the necessary retraining and reskilling for future job requirements. Discuss examples of successful programs while analyzing the public sector and private sector roles in program delivery.

Global Perspectives on Automation

- The effects of automation operate differently in developing economic systems compared to developed economic systems. What difficulties emerge when developing nations attempt automation integration and what opportunities exist for them to harness automation for development purposes?
- Global automation shifts have transformed the worldwide labor market structure. The offshoring of jobs together with expanding global supply chains leads to modifications in domestic employment opportunities.

Long-Term Implications of Automation

- Human labor faces a changing fate because automation systems are expanding across society. A post-work society's cultural and psychological aspects together with its social implications become central to the discussion about its possibility.
- The implementation of automation systems offers opportunities to build sustainable economic frameworks. The automation industry holds the potential to enhance resource conservation while decreasing waste production and support the development of circular economic systems.

By using textbook knowledge with your personal experience and judgment and opinion you will make your post more attractive for discussion.

2.11 Projects

2.11.1 The Ethics of Bypassing Workplace IT Restrictions: An Undergraduate Inquiry

Abstract: The research will investigate the ethical considerations of employees skipping workplace IT restrictions based on a business newspaper's guide to circumventing corporate rules. The research project will dedicate a single semester to study how undergraduate students view this behavior while evaluating their moral reasoning and studying workplace implications.

Workplace policies and individual actions and technology integration create a rich ethical investigation area. The upcoming workforce members among undergraduate students currently link theoretical knowledge with future practical experience. This research will both help students grasp workplace ethics principles while offering perspectives on digital privacy attitudes and workplace autonomy and governance approaches among future professionals.

Research Questions: Undergraduate students exhibit what understanding of ethical conduct when employees bypass workplace IT restrictions. What are the perceived justifications or motivations for such behavior among this demographic? Do students understand the dangers and outcomes which result from breaking policies?

Methodology: The project will adopt a qualitative method due to its limited duration within one semester.

During the first month, the research will draw from existing literature that explores digital ethics alongside workplace surveillance and IT policies.

Undergraduate students will participate in focus group discussions to explore their detailed opinions regarding this matter.

Students will learn about two fundamental ethical theories: deontological and consequentialist ethics, which they will use to evaluate the behavior during classroom discussions.

Expected Outcomes: The research will uncover preliminary findings about the moral perspectives of undergraduates toward technological workplace practices and their professional conduct as future employees.

Key outcomes will include:

The report will compile the moral beliefs of the participating students. This research will show the difference between the ethical reasoning methods students use versus the actual workplace rules and their resulting consequences. The project will generate educational recommendations for faculty members to teach students about ethical technology practices in their future careers.

Significance:

The research findings about technology ethics from different generations will help organizations develop better workplace policies and educational programs. Students will gain essential self-reflection about their personal values and actions when they enter the professional sector.

Timeline:

- Week 1: Planning and Literature Review. Establish a project structure together with specific goals. Research will begin with the examination of violent media effects on school shootings and First Amendment limitations in this area.

- Week 2: Interviews and Survey Development. Selected experts will receive interviews. A survey instrument should be constructed and made ready for distribution purposes.
- Week 3: Data Collection. The survey distribution will start and end with data collection. The research team should begin data analysis as they receive each new submission.
- Week 4: Analysis and Reporting. Complete data analysis. The synthesized findings will be used to create the final report and the oral presentation. The research findings should guide the creation of recommendations.

Conclusion: The focused scope combined with an appropriate methodology makes this undergraduate research project a relevant exploration of ethical issues that future employees will encounter with technology at work. The study enables practical ethical theory application which develops students' academic abilities and professional readiness.

2.11.2 Digital Footprints and Workplace Conduct: A Case Study Analysis

Abstract: The modern world requires a digital connection between personal and professional life so that employees need to understand the boundaries between these spaces. This session aims to evaluate the ethical and professional implications of employee off-duty activities while evaluating their effect on company image through digital content distribution. The research examines video content distribution by employees through an example to evaluate employment termination procedures while establishing suitable disciplinary measures for involved parties.

The scenario establishes a basis to explore personal conduct in public spaces together with employee representation of a company and social media content sharing consequences. As future professionals undergraduate students need to understand how their workplace conduct affects their professional and social life and specifically how their employer identity influences their behaviour.

Research Questions: Is there an ethical basis for dismissing an employee because of their private life actions outside regular working hours? The level of company brand exposure determines the extent of misconduct severity. The study examines both employee and corporate legal exposure that could result from their actions. The paper investigates which disciplinary measures would be most suitable for all parties and how companies should make these decisions.

Methodology: This project uses an exclusive research approach because of its short duration.

The analysis of the provided scenario will be examined thoroughly by students through an evaluation of the setting and business reputation effects. A limited examination of academic studies will focus on digital ethics alongside employment laws and workplace policies about employee activities beyond work hours.

Survey: A brief survey of student opinions on the scenario will be conducted to gather a range of perspectives on the appropriate disciplinary actions.

Role-Playing Debate: Students will participate in a moderated discussion that includes different roles such as company leadership and the staff members and community members to present arguments for and against the termination.

Expected Outcomes: The project is expected to produce:

A thorough report about the ethical and professional and legal elements of the situation under investigation. Survey results summarize the varied student perspectives about the appropriate disciplinary measures. The research together with class discussions will provide evidence for disciplinary measures that students should follow.

Significance: The project aims to make students aware of the practical consequences of their behavior in both physical and virtual environments and how these actions affect their professional life. This project will help organizations understand how to address such situations while upholding both their brand image and employee rights.

Timeline:

- Week 1: Planning and Literature Review. Create an overview of the project along with its specific goals. Start researching how violent media relates to school shootings while also examining First Amendment restrictions.
- Week 2: Interviews and Survey Development. Conduct interviews with selected experts. The team needs to create and implement a survey instrument for distribution to participants.
- Week 3: Data Collection. The surveys will be distributed for participants to respond. Analyze the incoming data in real-time.
- Week 4: Analysis and Reporting. Complete data analysis. The report and presentation will combine study results for final submission. Use research findings to establish recommendations that will be presented in the final report.

Conclusion: The project provides undergraduate students with an opportunity to address an important contemporary matter which links digital technology to personal actions and workplace conduct. The analysis of this scenario from different perspectives allows students to understand the various aspects of employee conduct and digital information sharing in contemporary workplaces. The knowledge and skills obtained from this project will serve students well when they enter the workforce because their online and offline actions produce long-term effects.

2.12 Exercises

These questions are designed to help students and readers reflect on the evolution of automation, its impact on job roles, and the challenges and considerations that arise from these technological advancements.

1. How did the introduction of computers change the scope of automation in the workplace?
2. What types of jobs are most susceptible to displacement by automation?
3. How has automation transformed existing job roles?
4. What skills have become more valuable in the job market as a result of automation?
5. What is essential for individuals to do to remain competitive in an automated world?
6. What is job polarization, and how is it related to automation?
7. What are some measures that can help mitigate job insecurity caused by automation?
8. What is one of the most immediate benefits of technology in the workplace?
9. How has technology improved communication and collaboration among employees?
10. Name a way in which technology enhances security in the workplace.
11. How can technology contribute to cost savings in the workplace?
12. How does technology foster innovation within an organization?
13. What is a potential risk associated with over-reliance on technology?

14. Why might some employees resist the adoption of new technology in the workplace?
15. What are three advantages of the gig economy for workers?
16. How has technology contributed to the growth of the gig economy?
17. What technological advancements have accelerated the rise of remote work?
18. List two benefits of remote work for individuals.
19. How is automation expected to transform jobs rather than eliminate them?
20. What are some concerns related to job displacement due to automation?
21. Why is upskilling important in the age of automation?
22. What skills are becoming increasingly important due to automation?
23. How can organizations and governments support upskilling in the workforce?
24. What role does lifelong learning play in the age of automation?
25. How has technology improved workplace communication?
26. How do social intranet platforms and enterprise social networks benefit employees?
27. What barriers has technology dismantled in workplace communication?
28. What types of activities can employers monitor with current technology?
29. What ethical dilemmas are associated with employee monitoring?
30. Why is transparency important in employee monitoring?
31. What should clear monitoring policies outline?
32. What is the dual nature of technology in the workplace?
33. What does separating work and personal systems typically involve?
34. Why might separating work and personal systems enhance a professional image?
35. What is a security concern associated with merging work and personal systems? Why?

Chapter

3

Freedom of Speech in Cyber Space

- The history of freedom of speech and how it applies to the internet
- Examples of how social media and online platforms have impacted free speech
- Challenges and controversies surrounding online content moderation
- The role of government in regulating online speech
- Anonymity: Shield or Threat to Free Speech?
- Net Neutrality: A Matter for Regulation or the Market?

3.1 The History of Freedom of Speech and its Application to the Internet: Navigating the Digital Frontier

The history of human rights shows that freedom of speech evolved through time as new technologies appeared. The internet has transformed how we interact and exchange data while creating obstacles and prospects for this essential human freedom. This session investigates freedom of speech development through time alongside its modern application to the digital world while analyzing current digital age implications as described in [36, 209, 58, 60, 190, 94].

The Foundation of Freedom of Speech: Human beings have fought for freedom of speech since ancient times because they wanted to express their opinions freely without facing persecution. Freedom of speech received official recognition through the Magna Carta and the United States First Amendment which established its position as a basic democratic right.

The Internet's Impact on Freedom of Speech: The internet revolutionized communication through its ability to spread information widely which both strengthened individual voices and enabled massive data distribution. Through its ability to make expression more accessible the internet allows people to express their thoughts and oppose traditional ideas and engage in public discussions.

Internet freedom faces multiple obstacles which affect the right to free expression. Internet access has opened new communication channels yet it introduced specific difficulties for free expression.

1. Moderation and Content Regulation: Due to the massive amount of information online it becomes challenging to control and regulate prohibited content. The challenge for both platforms and governments exists in balancing content moderation with the protection of free speech rights. The challenge lies in finding this equilibrium because it demands analysis of cultural factors as well as user protection measures alongside maintenance of multiple perspectives.
2. Disinformation and Fake News: Through the internet disinformation and fake news have become widespread issues that hinder effective public communication. False information spreads quickly across online platforms because of their anonymous nature and viral sharing capabilities which divides society while damaging news source credibility. Educating users and promoting media literacy are vital in combating these challenges.

3. Online Harassment and Hate Speech: The internet created new opportunities for hate speech as well as cyberbullying and harassment to spread. People feel more comfortable engaging in abusive behavior because online interactions create both physical distance and anonymity. The issue of how to protect free speech rights against harassment while safeguarding individuals continues to be a complex problem.
4. Government Surveillance and Censorship: The internet faces varying degrees of governmental oversight which generates widespread worry about state surveillance and information restriction. The discussion about how to balance national security needs with free expression rights continues to be debated. Free speech requires protection from governmental practices which need both privacy safeguards and complete transparency.

Online freedom of speech needs sustained collective support from governments and technology companies and civil society and individual citizens.

1. Internet Governance and Legislation: Effective protection of freedom of speech demands legal frameworks that clearly define limits while respecting basic rights of free expression. Governments must collaborate to create protection systems for user rights and systems to address abuses.
2. Empowering Users and Media Literacy: The ability to recognize trustworthy information depends on the development of critical thinking abilities and media literacy skills by users. The negative effects of online speech can be reduced through education that teaches users how to behave responsibly online and practice digital citizenship while raising their awareness about potential dangers.
3. Transparent Content Moderation Policies: Technology companies need to create clear content moderation policies that provide both transparent guidelines and responsible accountability mechanisms and methods for users to appeal decisions. Decision-making transparency combined with active user involvement creates trust which develops into a welcoming digital space.

In conclusion, the internet has revolutionized the freedom of speech landscape by giving people new opportunities to express themselves through massive public discourse participation. The internet creates special challenges that need continuous conversation and community collaboration to manage. The protection of online free speech requires finding equilibrium between open communication freedom and managing content moderation together with disinformation and harassment and government oversight issues. Our goal should be to defend free speech principles while solving these challenges in order to create digital spaces which support inclusivity alongside diverse perspectives and unrestricted idea sharing.

3.2 The Impact of Social Media and Online Platforms on Free Speech: Empowering Voices in a Digital Age

The evolution of digital communication systems has enabled people to participate in public discourse through unprecedented channels for self-expression. The consequences of social media and online platforms on free speech exist in multiple layers of complexity. The session evaluates the current role of social media and online platforms in free speech while examining both beneficial and adverse effects during the digital era [36, 209, 60, 190, 94, 127, 151, 171, 204, 170, 221].

The Empowerment of Voices:

1. Amplication and Accessibility: Social media platforms enable users to broadcast their thoughts along with their personal narratives to a broad audience. The platform enables marginalized groups together with grassroots movements to broadcast their messages to broad audiences.

2. Global Dialogue and Connection: These platforms have established worldwide connections that have eliminated physical distance limitations. People from different cultural backgrounds can participate in dialogue to build cross-cultural understanding which promotes the exchange of ideas between different cultural groups. People connect digitally to create a stronger human rights dialogue about social justice and many political and cultural matters.
3. Citizen Journalism and Real-Time Reporting: Social media tools have allowed ordinary citizens to function as journalists who report on current events in real time. Users can create content that provides information about injustices and institutional responsibility through their firsthand accounts and real-time broadcasting and user-generated materials.

Challenges and Concerns:

1. Moderation and Content Control: Social media companies must handle content moderation by finding the right middle ground which safeguards free speech rights while blocking dangerous or prohibited content. The process of content removal together with algorithmic curation and community guidelines generates problems regarding transparency and consistency as well as potential biases in decision making.
2. Disinformation and Manipulation: The rapid dissemination capabilities of social media allow false information and fabricated news and propaganda to spread easily across the platform. Through manipulation of information and exploitation of weaknesses malicious actors spread misleading narratives which lead to public opinion shifts while damaging reliable information sources.
3. Online Harassment and Cyberbullying: Social media users experience the risk of online harassment as well as cyberbullying which prevents them from expressing themselves freely. People who use social media platforms experience both opportunities for self-expression and the danger of being targeted by online harassment and cyberbullying because of their platform presence.
4. Echo Chambers and Filter Bubbles: Online communities today experience both information segregation in echo chambers and content filtering through filter bubbles which shows users only similar content. Through algorithms and content feeds people tend to see mostly information that agrees with their current beliefs which results in limited exposure to other perspectives and hampers both constructive dialogue and the exchange of differing viewpoints.
5. Government Surveillance and Censorship: Many governments throughout the world attempt to regulate online platforms by enforcing monitoring operations and blocking content and shutting down internet services to limit free speech. Through these actions governments succeed in suppressing dissenting voices and restrict information access and idea circulation.

Promoting Responsible Digital Spaces:

1. Transparent Content Moderation: Social media companies need to establish clear moderation policies and guidelines that provide users with information about content rules and consistent enforcement methods and appeal channels. The transparency of moderation practices helps users trust the platform and express themselves responsibly while safeguarding their freedom to express themselves without excessive censorship.
2. Media Literacy and Digital Education: Digital education and media literacy education should be promoted because they prepare individuals to analyze information effectively while using the digital environment. The active participation of users in source evaluation and misinformation detection together with responsible engagement practices helps reduce the adverse effects of manipulated information.

3. User Empowerment and Digital Citizenship: User involvement as digital citizens represents a fundamental requirement for online communities. Users must learn about ethical behavior online together with methods to conduct respectful dialogues while understanding their influence on others in the digital world.
4. Advocacy for Open Internet: Open Internet advocacy works to protect free speech through maintaining an internet environment that supports net neutrality principles and equal access while fighting against government surveillance and control.

In conclusion, social media platforms together with online platforms have created substantial changes to free speech by enabling people to express themselves and participate in public discussions. The platforms have enabled people to express themselves while creating worldwide connections but they face issues with content moderation and disinformation and harassment and government control that need careful attention and proper governance. Social media platforms will maintain their role as free speech spaces by addressing digital citizenship concerns and promoting transparency which will create an environment for responsible online behavior.

3.3 Challenges and Controversies Surrounding Online Content Moderation

Online content moderation has emerged as a crucial function for digital platforms which strive to create secure spaces that attract users. This mission exists in an environment full of obstacles and conflicting viewpoints. The session will investigate the multifaceted problems that occur during online content moderation while examining the challenges platforms encounter and the disputes that develop in the process [77, 196, 199, 213].

1. Scale and Volume: Content moderation faces its greatest difficulty because of the massive quantity of user-generated material produced daily. Online platforms deal with billions of users while managing an endless flow of user-generated content which makes it extremely difficult to achieve effective monitoring and review. Platforms need to create advanced algorithms together with an extensive team of moderators to manage their massive content load.
2. Context and Cultural Sensitivity: The complexity of content moderation increases when moderators must understand cultural factors and contextual elements. Acceptable content in one cultural setting or context might prove offensive when viewed through another cultural lens. Platforms need to develop competencies to recognize cultural differences because misinterpretations and adverse reactions can occur otherwise.
3. Automation and Human Oversight: Most platforms utilize automated systems to help with content moderation tasks. The algorithms used in content moderation systems demonstrate imperfect performance which leads to incorrect positive or negative results. Human oversight serves as an ongoing requirement to prevent incorrect content removals while avoiding unintentional support of dangerous material. Maintaining proper automation and human involvement represents a major ongoing difficulty.
4. Freedom of Speech and Censorship: The practice of content moderation creates fundamental problems regarding free speech rights and censorship practices. Users require protection from hate speech and harassment and misinformation but responsible moderation needs to avoid restrictions on free expression. Platforms must achieve user protection through methods that honor diverse perspectives while maintaining user safety.
5. Political Bias and Selective Enforcement: Most content moderation controversies stem from allegations that platforms demonstrate political bias through their selective enforcement practices.

The way different political and ideological groups view treatment varies according to their individual perspectives. Platforms struggle to achieve unbiased moderation while being open about their processes to prevent discrimination against particular viewpoints.

6. Transparency and Accountability: Content moderation depends heavily on transparency for users to trust the process and experience fairness. The public together with users require precise guidelines and regular application of these guidelines. The public frequently questions platforms about their lack of openness regarding their moderation policies and decision-making processes. The requirement for enhanced accountability and open practices persists as an ongoing issue.
7. Decisions on Newsworthy Content: Content moderation decisions about newsworthy events along with public figures frequently generate intense public debate. The distinction between appropriate discussions and possible dangerous or incorrect information continues to grow more challenging with each passing day. Platforms receive intense criticism and public scrutiny when they manage this content because people often believe these platforms exhibit political bias in their decision-making.
8. User Privacy and Data Protection: User privacy and data protection issues arise during content moderation procedures. The moderation procedures require the collection and examination of user information to detect potential violations. Platforms need to find an equilibrium point between moderation effectiveness and privacy protection because they must ensure data management practices meet legal requirements along with ethical standards.

In conclusion, online content moderation presents multiple difficulties that include managing large volumes of data while respecting cultural differences and protecting free expression and political neutrality. Platforms need to regularly update their moderation approaches to handle current challenges while they uphold visibility and user data protection protocols. The digital environment requires continuous dialogue between platforms and their users and regulators to resolve both complex issues and controversies that emerge from online content moderation.

3.4 The Role of Government in Regulating Online Speech

Online speech regulation has become a highly complex issue due to internet and social media expansion. The management of freedom of expression alongside harmful content presents multiple difficulties to authorities. The discussion will examine government supervision of online speech while discussing advantages together with disadvantages and important factors to consider [83, 94, 106, 86, 117].

1. Government oversight of online speech regulation functions as a key component for protecting users from harm and maintaining their safety online. Through laws which fight hate speech and harassment as well as cyberbullying and illegal activities governments establish a legal framework that creates a safer online environment.
2. Public discourse and democratic processes experience negative impacts from the extensive presence of misinformation and disinformation across the internet. Public health emergencies and election processes and national security need specific attention from governments when they fight false information. Governments must implement measures to combat disinformation because these measures defend information integrity and protect public interests.
3. Online speech regulation requires finding the right equilibrium between safeguarding expression rights and reducing harmful conduct. Governments need to handle their regulatory actions with care because excessive limitations could limit both free speech and the unimpeded sharing of concepts. Governments need to create specific legal structures and guidelines which determine what speech remains protected versus what content constitutes harm.

4. Online speech regulation by governments creates worries regarding both political biases and censorship practices. Excessive government regulations provide authorities with too much control over information distribution and enable them to suppress dissident opinions. A proper equilibrium requires both transparency and accountability measures along with strong protection systems to prevent regulation from infringing on democratic principles or violating individual freedoms.
5. Online speech regulation requires governments to work together with technology companies for effective implementation. The combination of government agencies with platforms enables responsible policy development as well as content moderation rules and reporting systems. Through cooperative work between governments and technology companies both groups can handle problems better while ensuring their regulatory frameworks match modern technology developments.
6. The oversight of online content extends beyond national frontiers since it exists in a global online space. The worldwide reach of the internet creates three main regulatory hurdles which include questions about legal authority and enforcement capabilities between borders as well as different national laws. Governments need to handle these complicated issues by working internationally to manage online speech challenges effectively and protect both sovereignty and cultural uniqueness.
7. Online speech regulation by government agencies requires the highest protection for user privacy combined with fundamental rights. The protection of individual rights such as privacy and freedom of expression must accompany online regulation measures to prevent excessive restrictions. Governments need to determine privacy effects from their actions and verify that all regulatory steps are necessary and proportional.

In conclusion, online speech regulation by government entities presents a multifaceted problem that requires careful evaluation. Governments must protect their citizens from harm and create safe digital spaces but they must achieve this through methods that respect democratic values and individual rights to expression and avoid censorship or political bias. A balanced method for regulating online content requires teamwork with technology companies alongside international partnerships and thorough evaluation of privacy and user rights. Governments who handle these challenges effectively will establish an online environment which supports open dialogue alongside protection of user well-being and harm reduction.

3.5 Anonymity: Shield or Threat to Free Speech?

Democratic societies recognize free speech as their fundamental basic right. Free speech remains complicated because individuals must confront the problem of speech anonymity. Public discourse experiences multiple conflicting effects from anonymity which requires thorough investigation.

Anonymity serves as an essential tool for free expression since it eliminates identification requirements. Many individuals find the ability to stay unidentified allows them to share minority opinions as well as dissenting and unpopular viewpoints both online and in print and at public protests. People who express controversial views through anonymity gain protection from harassment and retaliatory actions. The practice of being unidentified serves whistleblowers and activists to reveal illegal activities without risk of punishment. Anonymity-enabled applications allow people to communicate through social boundaries as documented by [60, 213, 30, 235, 139].

The lack of identifiable information diminishes personal responsibility which might lead people to produce dangerous statements. The spread of hate speech alongside misinformation and cyberbullying and online mobs becomes more probable when speakers face no accountability through identification. Anonymous users can infiltrate social communities to create division among members. Repressive governments enforce restrictions on anonymity because it helps them control dissent.

The anonymity system of Reddit enables free speech but the platform faces multiple negative consequences. Protection measures against dangerous speech need to maintain equilibrium with the defense of free expression rights. The push for verified social media identities demonstrates widespread worries about anonymity yet verification methods create privacy risks and exclude certain groups.

Anonymity creates a dual impact on free speech by either increasing its range or reducing its accessibility. The internet's development forces democratic societies to establish a proper equilibrium between supporting beneficial anonymity and controlling its possible dangers. Pseudonymity represents one of several approaches which demonstrates how to achieve middle ground solutions. Open communication requires freedom of speech to function because it needs honest discourse as well as accountable dialogue to enable human progress and connections.

In summary, open communication faces two risks from anonymity because it offers beneficial opportunities and harmful consequences. Anonymity powers speakers effectively when managed through appropriate restrictions based on ethical norms during reasoned discussions. No solutions exist in perfect form yet exploring these complicated questions with open discussion among groups may lead to balanced outcomes.

3.6 Net Neutrality: A Matter for Regulation or the Market?

For years net neutrality has been a widely debated and discussed concept that defines how Internet service providers (ISPs) should treat data on the Internet. The core issue of the net neutrality debate is the debate on how best to preserve a free and open internet and ensure fair access to all users. The main issue is whether net neutrality should be guaranteed by the government or left to market forces to decide. This session presents both the case for regulatory intervention and the case for market-driven solutions in the net neutrality debate [12, 58, 188, 232].

Regulations: Safeguarding Openness and Fairness

- Protection Against Discrimination: Net neutrality regulations prevent ISPs from discriminating against specific types of content, applications or services. In the absence of such rules, there is the danger of ISPs favoring their own services or those of their partners, thereby limiting competition.
- Protection of Freedom of Speech: The regulations protect freedom of speech by prohibiting ISPs from blocking or throttling content based on political or ideological considerations. This is important to ensure that all voices can be heard in the digital age.
- Promoting Innovation: Net neutrality rules ensure that all online businesses have an equal chance. In the absence of such regulations, large companies with resources to pay for prioritized access can be expected to limit innovation and competition.
- Consumer Protection: These regulations protect consumers by making sure that they receive the quality of service that they pay for and not experience artificial slowdowns or disruptions of specific content.
- Transparency and Accountability: The regulatory frameworks ensure that ISPs are transparent about their practices and this allows consumers to choose their internet service providers wisely. This transparency fosters accountability.

Market: Encouraging Competition and Innovation

- Market Competition: Some claim that net neutrality regulations can deter market competition as they may discourage investment in infrastructure. Competition between ISPs allows them to compete on the basis of the quality of service and price, which may result in better choices for consumers.

- Flexible Solutions: The market is better placed to respond to the changing tastes of consumers faster than regulations can. As technology advances, market forces can bring about new solutions to internet access and content delivery.
- Avoiding Regulatory Overreach: Net neutrality regulations have been criticized for being too much government intervention which may have negative consequences and create bureaucratic red tape.
- Differentiated Services: In a market driven approach, ISPs can provide different services that are determined by the consumers' needs. This could result in specialized packages that meet particular needs and thus increase user options.

Striking a Balance

The issue of net neutrality is very complex and has many aspects to it and both sides of the argument are correct. Finding the right balance between regulatory oversight and market forces is necessary. The regulations that are too stringent may hamper innovation and investment in the internet infrastructure, while a market with no regulations at all may result in discriminatory practices and consumer exploitation.

In reality, most countries have adopted a mixed model where some net neutrality principles are written into law and ISPs are allowed to offer tiered services. This is an attempt to provide a level playing field and allow for innovation and competition.

In conclusion, it is important to maintain the core principles of net neutrality: equal treatment of data, no blocking and no throttling, but at the same time adapt to the changing world of the Internet and the needs of users. This should be achieved through regulation, market forces or a mix of the two with the aim of preserving a free and open internet that is beneficial to all users and promotes innovation and competition.

Here are some examples of violations of net neutrality principles by internet service providers and governments:

1. Throttling: In 2014, it was disclosed that AT&T was throttling the speeds of its customers with unlimited data plans once they exceeded certain data thresholds. This slowing of selected traffic violates net neutrality.
2. Telus the Canadian telecom restricted workers who supported the union strike from accessing a particular website during 2005. Net neutrality defenders considered this action a violation of free speech principles.
3. In 2018 Verizon restricted the data speeds of California firefighters who used "unlimited" data plans before offering them paid premium data upgrades. The practice of prioritizing specific data sets directly violates neutrality principles.
4. The "Sponsored Data" platform of AT&T which allows customers to use data-free content from partner websites raises concerns about preferential treatment that breaks neutrality principles.
6. Bandwidth throttling: The internet service provider Vodafone Egypt was accused of throttling bandwidth to force users toward expensive service plans which violated neutrality principles.
7. The mid-2000s saw Madison River Communications alongside other US providers block competing Voice over IP services which violated FTC principles of neutrality.

These examples show that both corporate entities and governmental bodies fail to preserve internet traffic equality through their discriminatory treatment of online content. Free digital access becomes compromised by such violations.

3.7 Case Study 1

RESEARCH ARTICLE

Understanding international perceptions of the severity of harmful content online

Jialun Aaron Jiang*, Morgan Klaus Scheuerman, Casey Fiesler, Jed R. Brubaker

Department of Information Science, University of Colorado Boulder, Boulder, Colorado, United States of America

* aaron.jiang@colorado.edu

Abstract

Online social media platforms constantly struggle with harmful content such as misinformation and violence, but how to effectively moderate and prioritize such content for billions of global users with different backgrounds and values presents a challenge. Through an international survey with 1,696 internet users across 8 different countries across the world, this empirical study examines how international users perceive harmful content online and the similarities and differences in their perceptions. We found that across countries, the perceived severity consistently followed an exponential growth as the harmful content became more severe, but *what* harmful content were perceived as more or less severe varied significantly. Our results challenge platform content moderation's status quo of using a one-size-fits-all approach to govern international users, and provide guidance on how platforms may wish to prioritize and customize their moderation of harmful content.

Introduction

It is no secret that social media platforms have been plagued with harmful content—issues like misinformation and the incitement of violence. As a result, platforms actively moderate content to prevent users from harm and free their sites from abuse. However, as social media platforms have grown to a global scale, the people, and their associated values and behaviors, that platforms need to moderate have also become increasingly diverse [1]. As a result, moderating the enormous amount of content they generate becomes a significant challenge. Take Facebook, currently one of the largest social media platforms, which has 2.7 billion monthly active users worldwide and supports 111 languages. Despite its size, it has about 15,000 moderators [2] compared to billions of its users. It also has one set of rules, the Facebook Community Standards [3], guiding that moderation.

Facebook moderators need to enforce the Community Standards consistently, throughout the world—but how do we ensure consistency when the moderators and the people they moderate come from different backgrounds and cultures? For example, prior research has shown that the acceptability of sexualized nudity varies between cultures [4], and exiting news reports have also shown that translation of community guidelines did not extend to all languages [2]. When a moderator in Hyderabad needs to review content created by someone in Austin, as a result of globally distributed moderation [5], do they perceive the content and the governing rules consistently? Furthermore, as the moderation resources are limited compared to the extremely large amount of content that they need to handle, how should platforms prioritize moderation resources? Should that prioritization change in different parts of the world?

https://doi.org/10.1371/journal.pone.0256762 Used with CC 4.0 license

Prior research (e.g., [6, 7]) has suggested that *severity* can be a good heuristic to differentiate and prioritize different types of harmful content; violations that are more severe should be prioritized than those less severe. Severity, in this case, refers to the intensity of harm of violating content. From a severity prioritization perspective, a deep understanding of severity is necessary for actionable guidelines for moderators to make more informed moderation decisions, for social media platforms to more effectively prioritize moderation capacity, and for regular users to better understand online rules. A severity prioritization perspective can also guide the different levels of consequences of violating the rules.

Therefore, in this paper, we quantitatively investigate the severity of different types of harmful content through a survey with people from eight different countries in the world. In presenting our work, we first introduce prior work on social media content moderation. We then detail how we developed our survey, highlighting a novel method to measure participants' perceptions of the severity of identified types of harm: willingness to fine. We present each method of analysis alongside findings by research question. In presenting the findings, we answer the following research questions:

- How do people in different countries perceive the severity of harmful content?
- What are the similarities and differences between perceptions of severity in different countries?
- How do countries agree or disagree on the severity of different types of harmful content?
- What are some sensitive topics for specific countries?

Background: Content moderation on social media platforms

Social media platforms, like Facebook, Twitter, and YouTube, have become a central part of the social lives of billions of people across the world. While these platforms are not by themselves content producers, they are responsible for storing, organizing, and circulating a massive amount of user-generated content [5, 8]. Despite their claims of being impartial and their reluctance to regulate speech [1], many platforms are incentivized to moderate. They must not only meet legal and policy requirements, but also avoid losing users subject to malicious behaviors, protect their corporate image, placate advertisers who do not want to be associated with negative online communities, and honor their own institutional ethics [8, 9]. Through a case study of Reddit's ban on hate communities, prior research has shown that these moderation efforts can be effective in improving otherwise toxic online communities [10].

Platforms often govern users with two types of rules: terms of service and community guidelines. Terms of service serve a legal contract between the platform and the users; they spell out each party's obligations and liabilities [1]. Jackson et al. have called for deeply integrating the role of policy into research, pointing out that policy is deeply intertwined with design and practice [11].

Community guidelines, on the other hand, often use plain language that explains platforms' expectations of proper user behavior. Compared to terms of service, users are more likely to read community guidelines and understand them when they have the need to reference platform rules [1]. Platforms often impose community guidelines that detail acceptable and unacceptable behaviors. They prohibit illegal behaviors, such as engaging in child exploitation and human trafficking, but they also police non-illegal upsetting content, like harassment and commercial spam that could drive users away. Gillespie argues that these rules are important for numerous reasons. On one hand, they contain abusive behaviors and set the norms of platforms. On the other, they also help construct an ecosystem of governance that smaller platforms may look to for guidance, sometimes borrowing their languages directly [12]. Larger platforms may also adjust rules and policies in relation to each other [8]. A recent example is Twitter's ban on political ads immediately following Facebook's

https://doi.org/10.1371/journal.pone.0256762

reluctance of removing or fact-checking them [13]. These rules, especially when they are different, reveal how platforms see themselves as the arbiters of cultural values.

As platforms grow to have millions, even billions of users who produce massive amounts of content, scale presents significant challenges to platform content moderation [5]. The challenge of scale is two-fold. First, the vast amount of content eliminates the possibility of *ex-ante*, or proactive moderation, where moderators review content before it can appear on the platform. However, almost all platforms have to resort to *ex-post* moderation, or reactive content moderation, which initially allows all content to be published without review. Moderators then remove or filter questionable content after the fact, often after it has been reported by users [5, 8]. However, this approach results in malicious content impacting users on the platform, even if it only stays up for a short period of time. Second, moderating an immense amount of content also requires a large human workforce. In 2019, Facebook alone had 15,000 full-time moderators across the world to combat malicious content, but this number is clearly outpaced by the billions of generated content they must review [14]. Some disturbing content therefore remains online for days, months, or even years due to the lack of moderation capacity among other complications in large scale content moderation [1]. Many malicious users also try to circumvent moderation such as by using lexical variations, which makes timely moderation even harder [15, 16]. Platforms have been using automated algorithms to assist with removing abusive content [17], but the accuracy of algorithmic moderation still has much to be desired.

As platforms become global, platform rules start to intersect with the various but granular local laws [18]. For example, the EU has rules against certain types of hate speech [19]. Such is also the case with holocaust denial. While many platforms commit to remove such content, it is only a legal requirement in certain countries including Austria and Germany [20]. At the same time, the massive amount of users from various cultures and backgrounds also bring a wide variety of norms and values, presenting another challenge to scaling up human moderation labor [1], especially for platforms with a single set of rules for users across the world. Popular news has been rife with controversial moderation decisions. For example, when Facebook moderators removed the Pulitzer Prize-winning Vietnam war photo featuring children fleeing a napalm attack for child nudity [21], or when Facebook and Twitter seemed to exempt politicians spreading misinformation from removal or fact-checking. The debates about how to handle moderating contentious topics, like prize winning photos and misinformation, reflect a clash of different backgrounds, purposes, and values. Recent work has argued for a "constitutional layer" in digital institutions to make changes that are sensitive to local contexts [22]. Our research speaks to potentially highlighting this clash of values by examining how people from different cultures and countries perceive rule violations.

Regardless of a proactive or reactive approach, human moderators need to review significant amount of content, whether violating or not, 24/7. To meet such high moderation demand, many platforms resorted to outsource content moderators [23]. In her work about commercial content moderators [5], Roberts pointed out that the factory-like routine of content moderation work has led many moderators to burn out. The constant viewing of disturbing and traumatizing content takes a heavy emotional toll on the moderators, who are reluctant to discuss their work with their family and friends to avoid burdening them [14].

Methods

In this study, we deployed a survey globally to capture regular internet users' perceptions of the severity of harmful online content. In this section, we first detail how we implemented the survey, including survey question development, participant recruitment, and language translation. We then describe our data cleaning procedures in preparation for our analysis. The full survey instrument as well as participant response data are included in S2 File.

https://doi.org/10.1371/journal.pone.0256762

Survey development

Given that the goal of our study is to uncover culturally situated perspectives of harmful content online across the globe, we designed the survey with two primary goals: to be both (1) generalizable across platforms, and (2) representing a variety of countries.

Survey questions. To ensure the generalizability of survey results to different platform community standards, we asked the participants to rate each type of harmful content identified in Jiang et al.'s cross-platform content analysis [24]. Because participants may not have the same understanding about what each type of harmful content was, we asked participants to consider and rate each type of harmful content within the context of a hypothetical scenario, where they needed to imagine someone violated a particular rule. For example, participants rated the severity of "revenge porn" by responding to a hypothetical scenario where they see someone posted photos of revenge porn. We also described the scenarios as if they happened on Facebook, specifically, for two reasons. First, Jiang et al.'s analysis showed that Facebook's Community Standards was the only one that covered every type of harmful content identified. Second, concretely grounding the scenarios to a single platform could also mitigate the potential platform-related variability in perceptions of the same harmful content (e.g., people may perceive harassment differently on Facebook than on Twitter).

We employed a novel method for assessing the perceived severity of harmful content, given the limitations of common survey measurement instruments. The major limitation we needed to address when designing our survey was that inherently bounding the concept of "severity" could cognitively bias participant responses. Further, it would inherently bound severity to a linear scale—rather than a potentially exponential one. While Likert scales are a type of ordinal scale are common for quantifying respondents' opinions in survey design, using a bounded scale that typically has only five or seven options would also limit our ability to capture the nuanced perceptions between 66 different types of harmful content, and to model an accurate relationship of severity. Therefore, we asked participants to directly report how they perceived the severity of harmful content through free-text numerical values, following an established alternative to Likert scales in psychology research [25, 26]. In other words, participants could freely input a number as their answer. A free-text input would both: (1) not limit participants to a bounded linear scale in their assessments of severity; and (2) allow us to examine the rate at which severity increased (e.g., linearly or exponentially).

As the severity of harmful content can be an highly contextual and vague concept to regular users, in the survey we operationalized severity with two more concrete concepts: punishment and urgency. The construct of *punishment* represents the negative consequences a user believe is deserved for a certain type of harmful content. Degrees of punishment is rooted in the concept of proportionality in criminal justice literature: the more severe the harm is, the graver the associated punishment should be [27]. The construct of *urgency* addresses platforms' problem of prioritization, which in practice guides the operation of many emergency response systems [28]: the more severe the harm is, the sooner it should be taken care of.

To measure the two constructs of punishment and urgency, inspired by the willingness-to-pay measurement in economics research [29], we asked participants to respond with the amount of money that: (a) they would fine the person for posting the harmful content; and (b) the social media company should spend to remove the harmful content immediately, over all other types of harmful content. In answering these two questions, participants could input any number as low as zero (which indicated that they believed the content was not harmful), and as high as they would like. To compare the results between the free-text numeric measurements and traditional Likert-scale measurements, we also asked participants to rate the severity on a Likert scale that measured how upset the participant would be upon seeing the harmful content.

https://doi.org/10.1371/journal.pone.0256762

As such, participants answered the same three questions for all scenarios that they saw. Below, we show an example of both the scenario and accompanying survey questions a participant would see (in this specific example, we ask about "hate speech: dehumanization"):

> Imagine you saw:
>
> A post that says **people from a certain country are insects** on Facebook.
>
> - How much money, if any, would you fine the person who posted this content? Please indicate your answer in {local currency}. You only need to enter a number. (Punishment measurement, free-text numeric)
> - How much money, if any, do you think Facebook should spend to remove this content immediately over other types of content? Please indicate your answer in local currency. You only need to enter a number. (Urgency measurement, free-text numeric)
> - How upsetting is this content to you, if at all? (Upsettingness measurement, Likert scale: extremely, very, somewhat, a little, not at all upsetting)

Recruitment. For the survey results to be representative of multiple countries, we strategically recruited survey participants from eight different countries through Qualtrics panels, based on published statistics of countries with the most Facebook users [30]. While this is certainly not a comprehensive list of countries for the overall social media landscape, we feel that this sample has the level of diversity needed for the purpose of our research. We recruited 1,696 participants in total, with approximately 200 participants per country. This sample size was the result of budget constraints, having approximately equal representation per country, and having a sample size with a reasonable margin of error. For example, India, which had the largest Facebook user base, had 269 million Facebook users as of October 2019. Therefore, a sample size of 200 could give a satisfactory 7% margin of error with a 95% confidence level.

The participation criteria were that the participant was over 18 years old and had used Facebook in the last 30 days. To get the most accurate sample of Facebook users, we did not specify that we were specifically looking for Facebook users. We asked participants to choose all the social media platforms that they used from a list of options, but they would only be qualified if they had chosen Facebook as an option. This sampling process was necessary as the survey scenarios were described in the context of Facebook. Following recommendations in survey methodology literature [31], we also had participants answer a commitment question in the beginning that asked whether they agreed to thoughtfully provide their best answers, and would be screened out if they could not promise to do so. We also collected demographic data including age range, gender, and education level at the end of the survey to prevent early drop-off.

Piloting and translation. We piloted the survey with 36 pilot participants to test the validity of our approach. All pilot participants were able to understand the questions correctly and provide reasonable answers. However, according to their feedback, asking about all 66 types of harm would make the survey overly lengthy, with a completion time of approximately 30 minutes. Therefore, to prevent fatigue and drop-off, in the final version of the survey, each participant saw a random 33 types, half of the 66 total types of harmful content. The randomization effectively made each type of harm receive approximately 100 ratings per country. While we acknowledge that the sample size could only provide limited statistical power to generalize to the population, we feel that the diversity of the respondents could still reveal useful insights.

https://doi.org/10.1371/journal.pone.0256762

Finally, since many parts of the world do not speak English as the dominant language, we also had the survey professionally translated into the corresponding dominant language spoken in each country. The {local currency} placeholder was also translated into the country's dominant currency. For example, U.S. participants would see "U.S. Dollars (USD)," while Turkish participants would see "Türk Lirası (TRY)" ("Turkish Lira" in the Turkish language). Table 3.1 shows the final list of countries where we deployed the survey, as well as the languages into which the survey was translated. The University of Colorado Boulder Institutional Review Board approved this study (Study Number 19–0695). Consent was given digitally, via online consent forms.

Table 3.1 Recruitment and translation details for each country in the study.

Country	Language	# Participants
Brazil	Portuguese	211
Egypt	Arabic	207
India	Hindi	213
Indonesia	Indonesian	217
The Philippines	Filipino	220
Turkey	Turkish	204
United States	English	215
Vietnam	Vietnamese	209

https://doi.org/10.1371/journal.pone.0256762.t001

Data cleaning

First, we had Qualtrics perform basic data quality checks to filter out obviously poor responses or non-responses. The checks included filtering out duplicate respondents, response time that was unreasonably long or short, and patterned responses, such as all zeros ("straight-liners"). We also conducted device IP address lookups to ensure the respondent was physically in the country we were intending to gather data for. we then conducted our own data cleaning, which involved two steps:

- **Winsorization:** Because some participants entered spuriously high values for highly severe harmful content (e.g., a 10^{17} fine), we winsorized the free-numeric responses at 95% level (i.e., capping all values in the top 5% at the value at the top 5% point) to reduce the effect of extreme outlier values.
- **Normalization:** Because the absolute monetary values varied depending on each participants' own conception of money, as well as the currency being used, these raw values were not comparable across different people or across countries. Therefore, using min-max scaling, a common normalization technique that scales raw values within a certain range, we normalized the monetary values to a maximum of 550,000, which is the median of the maximum values that the participants entered. Specifically, we used the following normalization formula:

$$X_{normalized} = \frac{(X - min(X)) \times 550000}{max(X) - min(X)} \quad (1)$$

where **x** was each participant's response to the 33 types of harm that they were presented, represented as a vector.

The normalization scaled all participants' response values to the same range of values, and therefore made them comparable with each other for our analysis. It also preserved the relative relationships between the numerical ratings for each participant.

How do people in different countries perceive the severity of harmful content?

We first explored how participants in each country viewed harmful content broadly, and how these perspectives compared with each other when juxtaposed. Because one of the goals of the study was to understand how different types of harmful content should be prioritized, we examined how each country prioritized harmful content, reflected in their rankings. Given that each country also quantitatively rated the severity of each type of harm, we also investigated the rate of growth of severity values, as a rank order is inherently linear and may not reflect the true relationships between category types.

To understand how participants in different countries perceive the severity of content types, we first calculated the mean punishment and urgency values for each type of harmful content across participants, for each country. We then calculated the overall severity value of each type of harm as the mean of its punishment and urgency values, for each country. In other words, if s_k is the severity value for harmful content k, then

$$s_k = \frac{(\overline{P}_k + \overline{U}_k)}{2} \tag{2}$$

where $\overline{P}_k$ is the mean punishment values for k across participants, and $\overline{U}_k$ is the mean urgency values for k across participants. We then plotted the overall severity values of each type of harmful content against its rank order by country.

As shown in Fig 3.1, we observed exponential growths in the severity of harmful content consistently across countries. To confirm the exponential growth, we conducted an exponential regression on each country's data using the statsmodels Python package. Note that we conducted regressions separately for each country, so the data used per regression was not nested in structure. Our randomization mechanism also precluded us from using a linear mixed model, as individually ranked harmful content are not likely to retain their rank positions on the full range of harmful content (i.e., hate speech with a rank position of 16 among 33 harmful content types is not likely to stay ranked the 16th when considered among 66 harmful content types). Table 3.2 shows the regression results using the exponential function $y = Ae^{kx}$, where y is the severity value for each type of harmful content, and x is the reverse rank order of each type of harmful content (i.e., 1 = lowest, 66 = highest, in the same order as the x-axis in Fig 3.1). Here, k determines the growth rate of severity: the higher k is, the faster the severity value grows as the type of harm becomes more severe. A represents the "starting point" of severity: the higher A is, the larger the severity value is for the least severe type of harm.

https://doi.org/10.1371/journal.pone.0256762

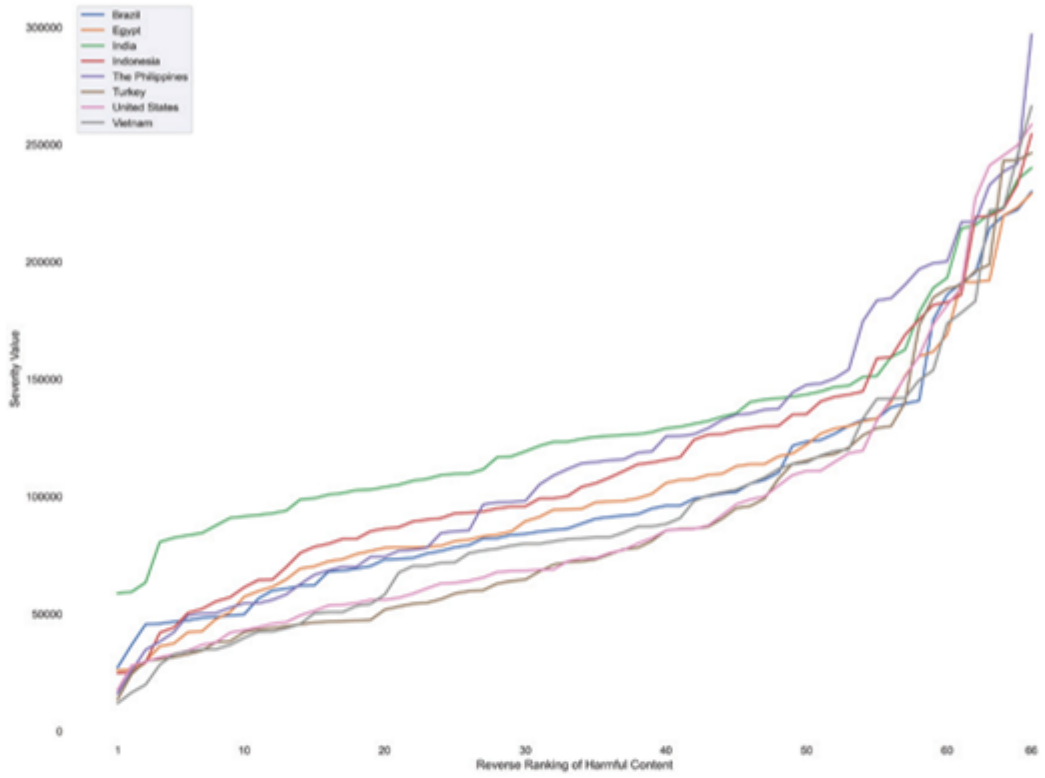


Fig 3.1 Plot of severity value vs. reverse severity rank order for each country under a free-text numeric measurement. Note that the same rank position may indicate different harmful content for different countries.

https://doi.org/10.1371/journal.pone.0256762.g001

As shown in Table 3.2, the regression analysis found that all parameters were statistically significant, thereby confirming the exponential of severity growth across countries. In our sample, Turkey had the highest rate of growth but the lowest "starting point" of severity, while India had the lowest growth rate but the highest "starting point." The consistent exponential growth means that, contrary to what a simple linear rank order would show, the "distances" between different types of harmful content were not equal—the "distance" between what was ranked in the first place (most severe) and the second place (second most severe) was much larger than that between the 65th place and the 66th place. While a rank order would imply a steady linear growth as a result of assigning consecutive integers, the exponential growth revealed that as the rank grew higher, not only did participants find the behavior more severe, it also gained severity more quickly.

We also tested the traditional Likert scale measurement as a comparison to the punishment and urgency free-response measurements, and we observed distinctly different relationships. Fig 3.2 shows a plot of the mean response of each type of harmful content versus its reverse rank order by country; the only difference with Fig 3.1 is the *y*-axis represents the Likert scale measurement, rather than the free-response measurement. As Fig 3.2 indicates, while the growth rates of severity still seemed consistent across countries, they were clearly closer to linear growth instead of exponential growth. A possible reason for the distinctive difference in growth rates is the inherent upper limit of Likert scales—in this study, the most severe possible option participants could select was "extremely upsetting," thereby disallowing free growth. By limiting the higher end of severity, the Likert scale may also flatten the nuanced relationships between different types of harmful content, which the free-response measurements were able to reveal. For example, under the Likert scale, the most severe harm in the U.S. was Child Exploitation Imagery (= 4.68), whose severity value was only 1.88x

Table 3.2 Exponential regression results of each country's data. Here, $p < 0.001$ for all parameters. See T1 in S1 File for the full set of regression results.

Country	*A*	A-	R^2
Brazil (BR)	40878.84	0.023	0.944
Egypt (EG)	39838.25	0.024	0.909
India (IN)	72922.22	0.015	0.918
Indonesia (ID)	43873.91	0.025	0.894
The Philippines (PH)	37959.63	0.028	0.931
Turkey (TR)	26118.68	0.03	0.95
United States (US)	28545.27	0.029	0.949
Vietnam (VN)	27099.36	0.031	0.914

https://doi.org/10.1371/journal.pone.0256762.t002

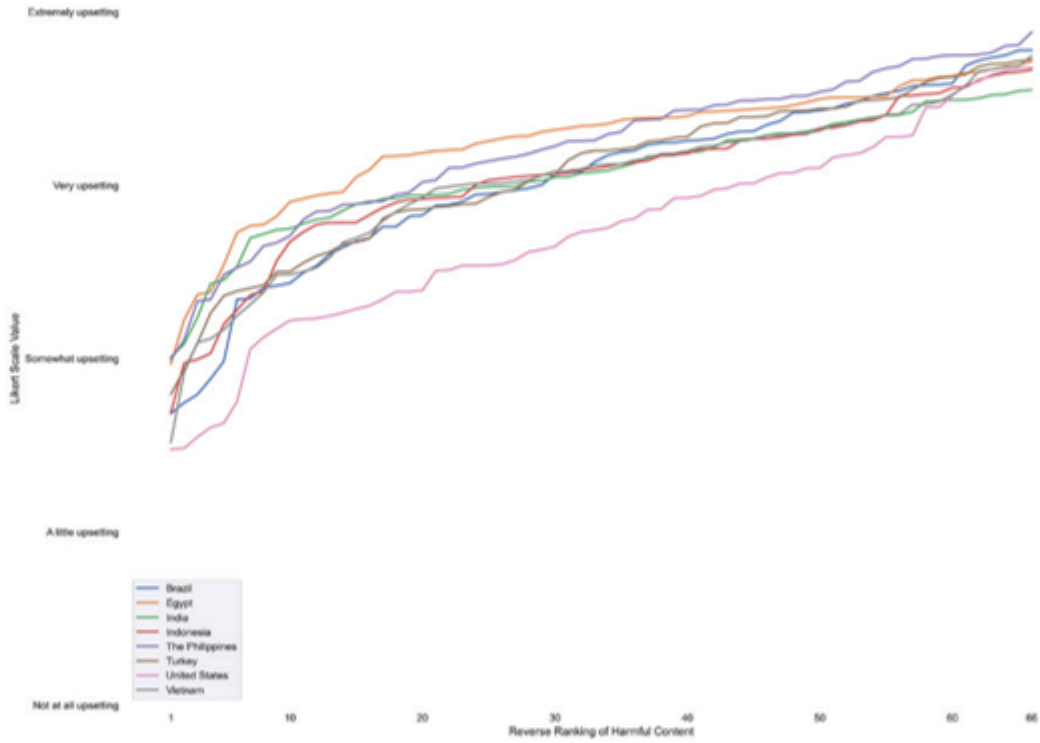


Fig 3.2 Plot of severity value vs. reverse severity rank order for each country under a Likert-scale measurement.

https://doi.org/10.1371/journal.pone.0256762.g002

that of the least severe harm, Engagement Abuse (= 2.48). However, the free-response measurement revealed that, without the arbitrary limit of the higher end, the severity of Child Exploitation Imagery (= 258632.86) was almost 15x that of Engagement Abuse (= 17453.15).

Overall, across the countries we examined, people's perceptions of harmful content's severity showed a distinctive pattern under a free-text numeric measurement: when the category of harm was ranked higher,

the perceived severity of the harm also grew exponentially. Furthermore, the commonly-used Likert scale measurement concealed the exponential growth by showing a linear growth instead, which would have underestimated the severity of many types of harmful content, especially those on the higher end of severity.

Despite the consistent exponential growth of severity, it is important to note that the specific harmful content in any rank position varied from country to country. In other words, participants in each country had their own ideas of how to rank different harmful content—what was the "worst" varied from country to country. We discuss these differences and similarities between countries in the following sections.

What are the similarities and differences between countries?

Overall, no two country rankings were the same. There was a lot more disagreement than agreement between countries, and there was always something that any two countries (strongly) disagreed on. However, we were able to observe clusters of countries within which individual countries were more likely to agree with each other.

To identify clusters of countries that are similar, we first calculated each country's ranking of harmful content, which is determined by the rank order of the average severity value s_k of each type of harmful content k across all participants in that country. For example, the most highly ranked harmful content in Vietnam would have the highest average severity value across all Vietnamese participants compared to the other types of harm. See T3 in S1 File for a complete list of country rankings.

We then performed principal component analysis (PCA) on these rankings of harmful content by country. PCA is a common technique to reduce dimensionality in a dataset [32], and was necessary because there were far more dimensions (i.e., the number of harmful content types) than data points (i.e., the number of countries), which would make for more expensive computation and less meaningful clusters [33]. We experimented with the number of components $n \in [2, 7]$ and found that when we projected the original 66-dimensional data onto $n = 7$ components, the result could explain full original variance (shown in Fig 3.3).

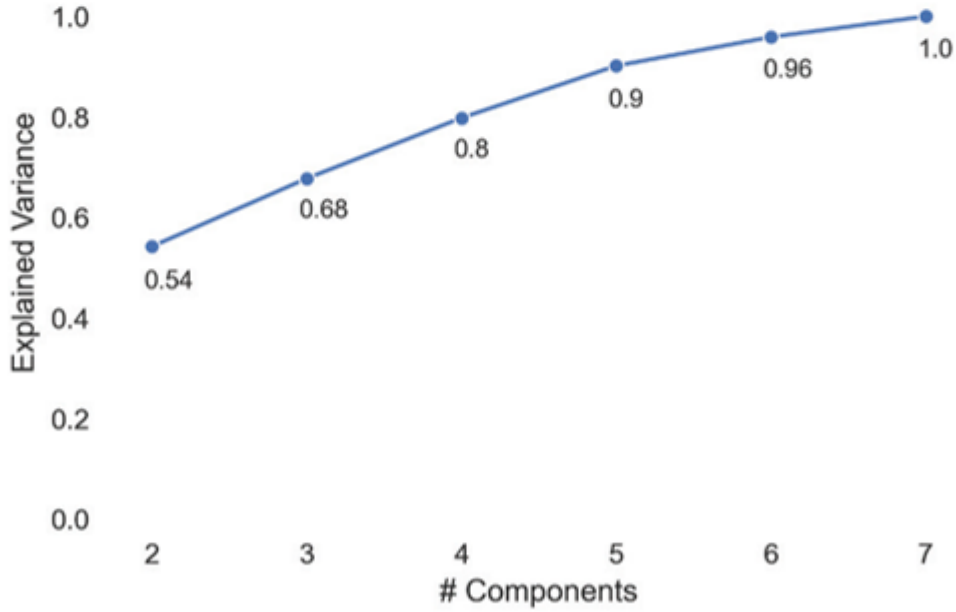


Fig 3.3 Plot of explained variance vs. number of principal components in PCA.

https://doi.org/10.1371/journal.pone.0256762.g003

After reducing the number of dimensions from 66 to 7 using PCA, we then used the Gaussian Mixture model to find clusters. We determined the number of clusters K by examining the average silhouette coefficient s for each possible K, which represents how well samples are clustered with samples that are similar to themselves [34]. The higher the silhouette coefficient s, the better the clustering is.

We tested the Gaussian Mixture models by experimenting with $K \in [2, 7]$, and found that the best clustering appeared at $K = 5$ ($s = .13$), before the silhouette coefficient started decreasing (shown in Fig 3.4).

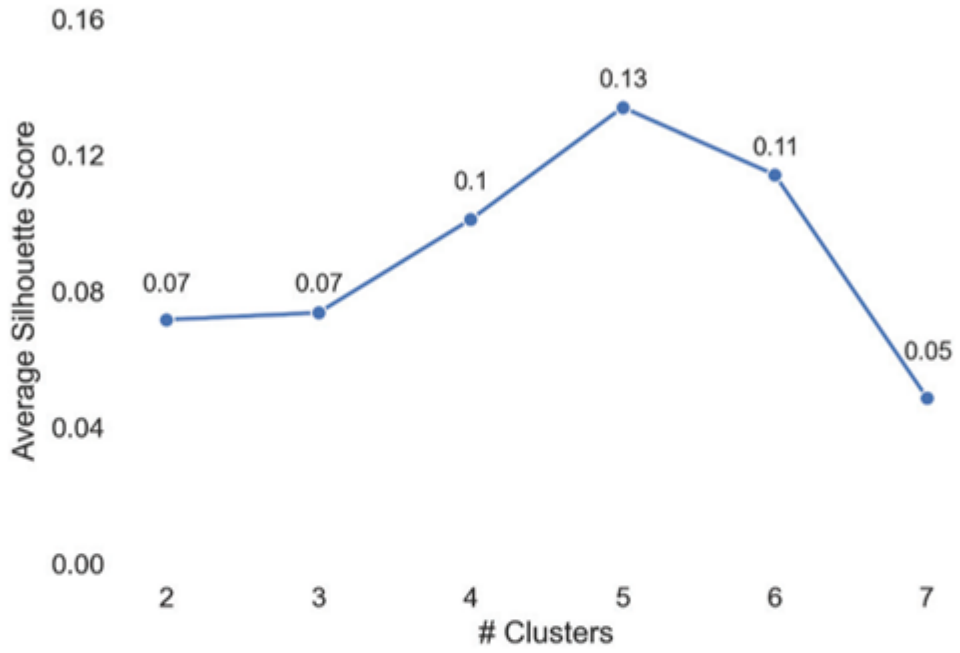


Fig 3.4 Plot of average silhouette score vs. number of clusters in Gaussian Mixture models.

https://doi.org/10.1371/journal.pone.0256762.g004

To confirm the robustness of our clustering, we also conducted a pairwise ranking correlation analysis; Table 3.3 shows the correlation results. The analysis also showed the same clustering, with within-cluster correlations consistently higher than 0.8. In other words, the rankings of countries within each cluster are highly correlated with each other, which corroborates with the Gaussian Mixture clustering result.

Table 3.3 Pairwise ranking correlation results for all countries.

	BR	US	IN	PH	EG	VN	ID	TR
BR	1	0.8835	0.6956	0.8138	0.6641	0.7310	0.6790	0.7660
US	0.8764	1	0.7491	0.8421	0.7407	0.6503	0.7017	0.7623
IN	0.6956	0.7491	1	0.891	0.8215	0.7606	0.7652	0.7609
PH	0.8138	0.8421	0.8910	1	0.8094	0.7586	0.8067	0.7943
EG	0.6641	0.7407	0.8215	0.8094	1	0.6336	0.6992	0.6729
VN	0.7310	0.6503	0.7606	0.7586	0.6336	1	0.8080	0.7557
ID	0.6790	0.7017	0.7652	0.8067	0.6992	0.8080	1	0.7697
TR	0.7660	0.7623	0.7609	0.7943	0.6729	0.7557	0.7697	1

https://doi.org/10.1371/journal.pone.0256762.g003

Each cluster of countries had its own set of harmful content that it perceived as more or less severe compared to other clusters, based on how they ranked these types of harm. Table 3.4 lists the types of harmful content perceived as more or less severe by each cluster.

Table 3.4 Country clusters and their ranking characteristics.

Cluster	Harmful Content Perceived as More Severe	Harmful Content Perceived as Less Severe
Cluster 1 Brazil (BR), United States (US)	Non-consensual Sexual Touching, Graphic Violence: Mutilated Humans, Hate Speech: Exclusion	Adult Non-consensual Intimate Imagery, Marijuana Sale, Drug Use, Vandalism
Cluster 2 India (IN), The Philippines (PH)	Distributing Virus, Sadism, Drug Use	Suicide Promotion, Creep Shots, Hate Org Coordination
Cluster 3 Egypt (EG)	Adult Non-consensual Intimate Imagery, Sexual Solicitation, Sexually Explicit Language, Coordinating Harm, Creep Shots, Criminal Group Coordination	Firearm Sale, Graphic Violence: Mutilated Humans, Self-injury Depiction, Harassment
Cluster 4 Vietnam (VN), Indonesia (ID)	Hate Speech: Dehumanization, False News, Marijuana Sale	Child Nudity, Minor Sexualization, Sadism
Cluster 5 Turkey (TR)	Eating Disorder Promotion, Eating Disorder Depiction, Harassment, Commercial Spam	Suicide Depiction, Criminal Group Coordination, Criminal Group Propaganda

https://doi.org/10.1371/journal.pone.0256762.t004

Overall, we found that each country cluster had a unique set of harmful content that they collectively perceived as more severe or less severe, and these sets rarely overlap between clusters. There are also types of harm that are ranked highly by one cluster, but low in another (e.g., "marijuana sale" has a high ranking in Cluster 4 but a low ranking in Cluster 1).

On what harmful content do countries agree and disagree?

While we have examined the patterns and differences between countries in terms of how they perceive severity, broadly, it is also valuable to conduct a similar investigation on the rankings of severity for different types of harm. Insights into the agreements and disagreements between individual types of content will reveal which types of harm may deserve more attention and nuanced treatment than others.

To examine agreements and disagreements, for each type of harmful content, we calculated standard deviations of all countries' rankings of that type of content, and plotted it against that type of content's overall sample-wide ranking (1 = highest), shown in Fig 3.5. Note that the overall sample-wide ranking here is a country-agnostic one, rather than the aggregate of by-country rankings. The reverse-U shaped curve in Fig 3.5 shows that countries agreed more on the most severe and the least severe harm, but disagreed more toward the middle (with "marijuana sale" having the largest standard deviation ($SD = 15.57$), and "engagement abuse" (which often appears as clickbaits) having the smallest standard deviation ($SD = 0.46$)).

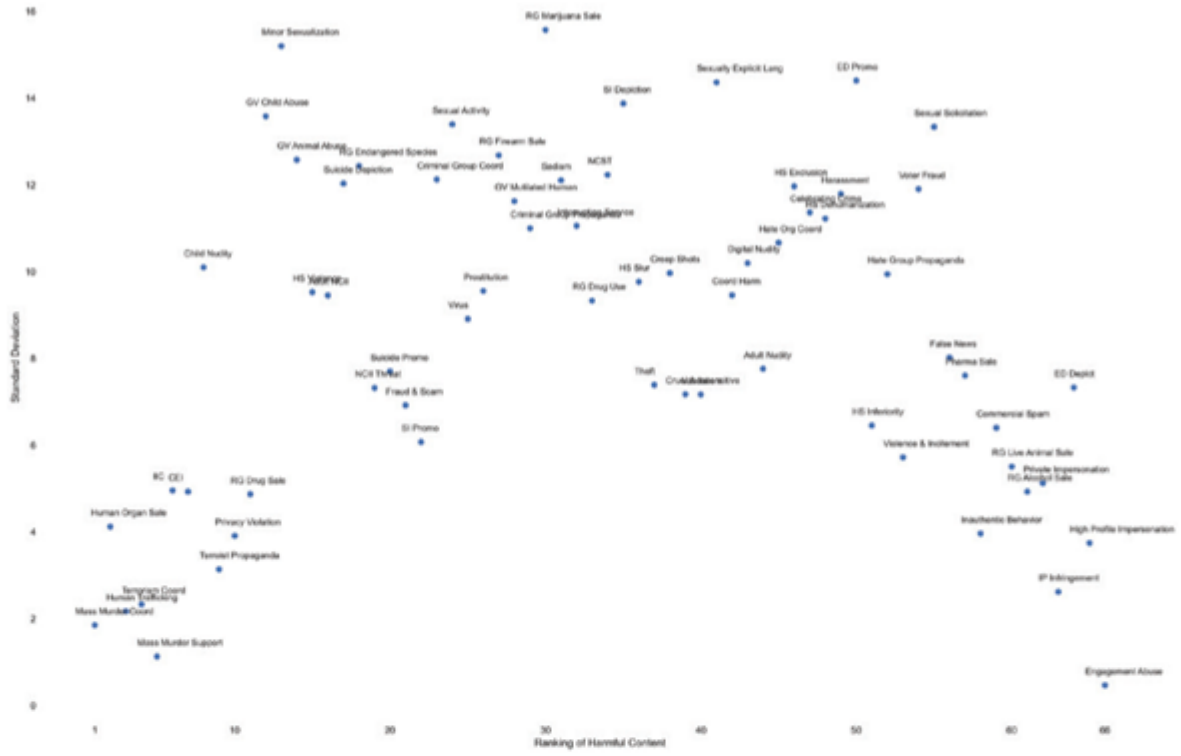


Fig 3.5 Harmful content's standard deviation vs. sample-wide ranking.

https://doi.org/10.1371/journal.pone.0256762.g005

While standard deviations can give us an general idea of how "spread out" countries are in their rankings of content types, we also examined the maximum extent that countries disagreed on a certain type of content by looking at "outlier" countries that ranked it most extremely. While it is common practice to exclude such outliers in data analysis, we chose to focus on them here because our unit of analysis is an entire country rather than an individual participant. Taking a closer look at the "outliers" will reveal valuable insights into the diversity of perceptions of harmful content. Therefore, we also measured the *max ranking differences*. Here, we define the max ranking difference ($\Delta rank$) within a type of content as its lowest ranking (i.e., highest in number) across countries less the highest ranking (i.e., lowest in number) of that content.

Our analysis showed that the largest $\Delta rank$ existed in "minor sexualization" and "self injury depiction," both with $\Delta rank = 45$. Similar to our standard deviation analysis, the smallest $\Delta rank$ appeared in "engagement abuse," with $\Delta rank = 1$. Similar to Figs 3.5 and 3.6 shows all 66 types of content's $\Delta rank$ plotted against their overall sample-wide severity ranking (1 = highest).

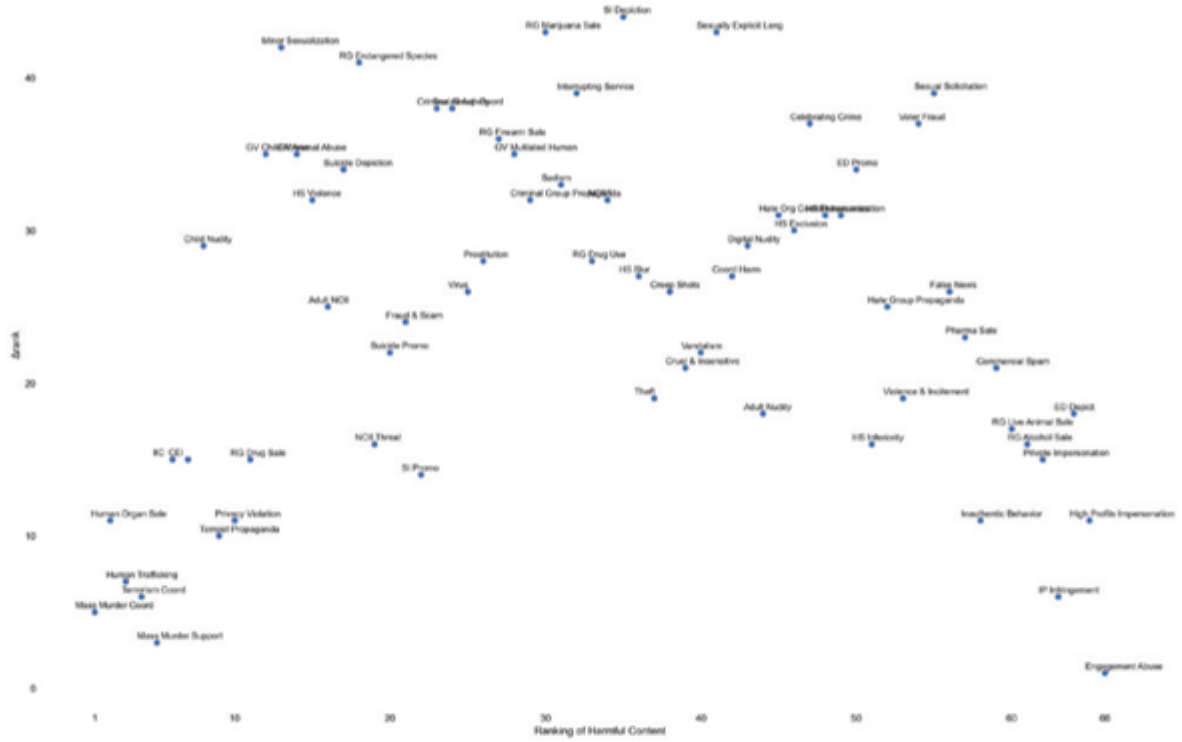


Fig 3.6 Harmful content's Δ*rank* vs. sample-wide ranking.

https://doi.org/10.1371/journal.pone.0256762.g006

To take a closer look at larger disagreements in Δ*rank*, we counted 18 types of content whose Δ*rank* are at least 33, or half of the total number of rank positions, as shown in Table 3.5.

Table 3.5 Types of harmful content that had max ranking differences (Δ*rank*) of at least 33, or half of the total number of rank positions.

Harmful Content	Δ*rank*
Minor Sexualization	45
Self Injury Depiction	45
Adult Sexual Activity	43
Regulated Goods: Marijuana Sale	43
Sexually Explicit Language	43
Regulated Goods: Endangered Species Sale	41
Graphic Violence: Mutilated Humans	39
Interrupting Platform Services	39
Voter Fraud	39
Sexual Solicitation	39
Criminal Group Coordination	38
Criminal Group Propaganda	38
Eating Disorder Promotion	38

Harmful Content	$\Delta rank$
Celebrating Crime	37
Graphic Violence: Animal Abuse	36
Regulated Goods: Firearm Sale	36
Graphic Violence: Child Abuse	35
Suicide Depiction	34

https://doi.org/10.1371/journal.pone.0256762.t005

While there is not a clear pattern in the kinds of content listed in Table 3.5, it is still worth noting that at least two countries would disagree on whether 19 types of harm, or nearly 30% of all identified types of harm, should be in the top or bottom half.

What are some country-specific sensitive topics?

The country-level and the content-level analyses in the previous two sections revealed that large disagreements in the perception of severity for different content. While there were clear clusters in perceptions, we found fewer patterns when examining individual types of content. Therefore, in an effort to further seek out patterns in the kinds of content that countries agreed or disagreed on, we employed a content analysis to qualitatively categorize individual types of content into higher-level topic categories.

First, the first two authors each independently categorized all 66 types of harmful content, with an eye toward achieving a reasonably small number of categories to more easily identify patterns. We also tried to follow how "common sense," rather than moderation experts, might interpret content types. For example, while the Facebook Community Standards [3] categorized "human organ sale" as part of "regulated goods," an average person may associate it with the black market or mass murder, and in our categorization we followed the latter rationale and categorized it under "Mass Scale Harm." We then came together to discuss and adjudicate differences, while iterating on the categories. While we acknowledge that our primarily U.S.-centered perspective is likely to have had an influence on our categorizations, we believe it is still a promising first step towards a preliminary understanding of the similarities and differences of individual content rankings in a multi-country context. We eventually agreed upon the following topic categories (see T2 in S1 File for a complete list of harmful content and their topic categories):

- Mass Scale Harm (e.g., Terrorism, Human Organ Sale)
- Vulnerable Groups (e.g., Child Exploitation Imagery, Child Abuse)
- Violence (e.g., Graphic Violence: Mutilated Humans, Sadism)
- Platform Abuse & Spam (e.g., Interrupting Platform Services, Engagement Abuse)
- Sexual Content (e.g., Sexual Activity, Sexually Explicit Language)
- Regulated Goods (e.g., Marijuana Sale, Drug Sale)
- Self-harm (e.g., Suicide Depiction, Eating Disorder Promotion)
- Financial harm (e.g., Fraud & Scam, Privacy Violation)
- Other Directed Harm (e.g., Theft, Vandalism)

Table 3.6 shows countries' median ranking of each topic category of content type. Overall, content involving "vulnerable groups" (e.g., children) are consistently the most severe across countries. The perceived high

severity is corroborated by the fact that there is at least one type of "vulnerable groups" harm in the top 6 (top 1%) of every country. Interestingly, while "mass scale harm" was only ranked third sample-wide and sometimes even lower in certain countries, it is consistently the majority category of every country's top 6. A possible reason for the relatively lower ranking of "mass scale harm," despite its concentration in the most severe harms, is that the category accounts for most *types* of harms (13 out of 66) and is thus more spread out in individual country rankings.

Table 3.6 Countries' median ranking of each type of harmful content. 1 = most severe; 66 = least severe. The World ranking is a country-agnostic ranking instead of an aggregation from individual country rankings.

Topic Category	Sample	PH	ID	US	IN	BR	VN	EG	TR
Vulnerable Groups	10	9	16	5.5	13	9	18	18.5	9
Financial Harm	15.5	15.5	19.5	15.5	16	19	15	22	17.5
Mass Scale Harm	23	18	25	27	20	19	23	15	42
Self-harm	28.5	33	23.5	30	36.5	31	40.5	31.5	28.5
Violence	31	46	42	23	43	27	40	43	41
Regulated Goods	31.5	35.5	25	42	31	46.5	23	48.5	31.5
Sexual Content	36	32	37.5	32	30.5	39.5	43.5	21.5	37.5
Other Directed Harm	43	41.5	39	46.5	43	38.5	32	43	36
Platform Abuse & Spam	60.5	57.5	56.5	59.5	57.5	59.5	56.5	61.5	61.5

https://doi.org/10.1371/journal.pone.0256762.t006

On the other hand, "platform abuse & spam" was consistently perceived as the least severe topic category of harmful content. Specifically, "engagement abuse," which refers to posting clickbait-like content, was consistently perceived as least severe (except in Vietnam and Egypt, where it was second to last.) Other "platform abuse & spam," including "intellectual property infringement" and "commercial spam" were also consistently in the bottom 10, showcasing they were not considered severe.

Discussion and conclusion

Overall, our findings show significant differences in perceptions of harmful content across different countries in the world. The general variability serves as evidence against social media platforms' current approach of using a single universal set of rules to regulate global users, which falsely implies that people view harmful content consistently across the world. The cross-country disagreement widely existed in different facets of our analysis—in different countries' rankings of content, in individual types of harmful, and in the higher level topics we used to categorize individual content types.

Despite the general variance, there were some patterns in how users perceived the severity of content. First, regardless of how each country ranked the content types, we found that the perceived severity followed the same rate of growth as the content ranking became increasingly worse—specifically, an exponential growth. Second, while there were no unified ways to describe how all countries ranked harmful content, there were country clusters where the rankings were similar, which suggests the possibility of moderation by country groups. Finally, in our analysis of the agreement and disagreement on the level of individual types of content, as well as high-level topics, the convergence toward the extreme ends suggests that people

had more consensus on the most severe and the least severe types of harm, but had less consensus toward the content in the middle.

Taken together, our findings have several implications for platform content moderation. First, the consistently exponential growth across all countries provides a principle for moderation prioritization in localized contexts—while what types of harm should be prioritized may vary from country to country, the underlying relationship between the types of harmful content may remain the same. The distinctive exponential growth also shows that using a simple ranking approach may be misleading because it implies a linear growth of severity and thus highly severe content may not receive the attention that it deserves.

Furthermore, our findings indicate that policymakers of social media platforms, who likely developed community guidelines from a predominantly Western (and particularly U.S.) point of view [1], may deprioritize the moderation of content that is perceived as highly severe in non-U.S. countries. However, we am not claiming that platform policy making should be in a "U.S. vs. elsewhere" fashion, because our evidence shows that perceptions of harmful content are complex and highly varying in countries outside the U.S. The risk of careless categorization is even more pronounced given that moderation resources and capacity are inevitably limited. Therefore, we argue that a promising direction for content moderation is to customize and localize by countries. At the same time, we also acknowledge that differences may still exist within countries, and more research is needed to uncover the specific nuances in any particular country. We also encourage the platform to and critically consider the value of adopting user feedback on content moderation, which we expand on later in this section.

Our analysis of country-specific sensitive topics revealed a similar pattern of consistent opinions about the most severe ("mass scale harm" and "vulnerable groups") and the least severe ("platform abuse" and certain kinds of "regulated goods") harm, but diverse opinions about the middle rankings. However, we present these findings as evidence for moderation approaches cautiously. While platforms may want to prioritize harmful content involving "mass scale harm" and "vulnerable groups" (as they may already be doing), it would be perilous to treat the least severe content as "not important," because our participants, regular social media users, may be unaware of the non-obvious harms that they could cause (e.g., "sale of prescription drugs" or "intellectual property infringement"). Further, the variance of perceptions of severity show the need for further research on understanding diversely perceived harms, as these harms are likely to be more complex. Therefore, simplistic moderation strategies for them may unintentionally privilege certain groups of people at the cost of others, a point echoed by Schoenebeck et al. in their study of different people's perceptions of justice models [35].

Methodological implications

On a methodological level, researchers and practitioners may find value in adopting the free-text numeric measurement method that we used within their own platforms or communities. By assigning unbounded values to individual types of harmful content, our method was able to reveal quantified relationships between values (e.g., one case of "child exploitation imagery" is equivalent to *x* cases of "commercial spam" in terms of severity) that the Likert scale could not. While reductive and not able to capture the full nuance of harmful content, these equivalence relationships can be a promising first step in guiding decisions in large-scale content moderation, including moderator resource allocation, proactive detection, and response prioritization, for different kinds of content, especially given the inherent limitation in human moderation capacity [1]. Researchers and practitioners can also conduct the same measurement survey with diverse groups of stakeholders, such as users, moderators, and policy experts, to gain a variety of perspectives. In the context of platform moderation, multi-stakeholder measurement may prove especially fruitful, because regular users who rarely encounter highly severe types of harm may not be able to reasonably estimate their impacts.

https://doi.org/10.1371/journal.pone.0256762

Ethical implications

Our study also has important ethical implications. While we have identified *what* types of behavior were sensitive and important for specific countries, we want to emphasize that we present no evidence for *why* they are important, which should be seen as a limitation of our study. We also intentionally avoided speculating possible reasons for the severity rankings we found. Irani et al. argued for a generative, rather than taxonomic, view of culture that positions people at the intersection of cultures [36]. Even though we conducted the study with some inherent regional taxonomies, geographically co-located individuals still experience and produce different, dynamic, and sometimes overlapping cultures. Therefore, reckless assumptions and speculations about "cultures" may not only be incorrect, but also highly dangerous, because they may impose harmful stereotypes and stigma on people. In the case of content moderation, the stakes are even higher because the resulting stigma may be associated with highly severe, possibly illegal behavior. Therefore, we caution against simplistic interpretations of these regionally different topics, such as assigning cultural stereotypes to these differences (e.g., "Vietnam users perceive 'creep shots' less severely than other countries because of their certain cultural values"). we urge researchers to conduct careful, deep-dive case studies into particular types of behavior to understand their full complexity.

Nevertheless, our research provides insights into the views of international users in platform policymaking, which has otherwise been centered around compliance with various laws and regulations. While it is unclear how exactly platform-level moderation policies were made, or whether user inputs were taken into account in the making of these policies, given that platforms currently do not have customized policies for different countries, it is reasonable to suspect that U.S. perspectives are weighed more heavily than those from other parts of the world, given the majority of large social media companies are headquartered in the U.S. However, the non-overlap and the opposite ranking of certain harms in our clustering analysis showed the necessity of treating different countries differently in creating rules and policies, rather than taking a one-size-fits-all approach that is likely to deprioritize harmful content that certain countries perceive as highly severe. Similarly, both our standard deviation analysis and the $\Delta rank$ analysis showed the necessity of customizing content moderation by country. A universal approach is likely to mistreat a nontrivial amount of harmful content by dismissing cross-country differences, and the level of harm across content types on which people heavily disagreed deserves greater attention, more in-depth research, and more nuanced treatment. Overall, our research complicates and challenges the current practice of blanket content policies for the whole world by revealing the diverse attitudes toward harmful content from users from different countries.

At the same time, it is important to note that universal approaches can sometimes be ideal, as over-customization may result in fragmented and inefficient content regulation policies. Our results also showed that certain issues such as child exploitation and the killing of humans indeed received cross-country consensus, which indicate blanket moderation may be advisable for similar issues. While there are no clear answers, platforms should carefully consider the trade-off between uniformity and customization, while keeping in mind the specific abuse they are dealing with, as well as current thorny debates such as online safety vs. freedom of speech.

The trade-off between universal and relativist content moderation policies and related ethical implications will be an everlasting one. While we have argued that regional differences are important considerations, it is also crucial to not assume that localized norms or laws are necessarily ethical. Our method does not speak to how to *resolve* conflicts between regional norms—rather, it helps *surface* and *highlight* these conflicts. Our method can be a way for platforms to understand how people in different countries quantify harmful or objectionable content, but in the end, platforms need to decide how best to reconcile regional differences, what norms align with platform values and priorities, and what they consider "right." For example, death sentences for LGBTQ people may be acceptable in some locations or cultures [37], but that does not justify acts of genocide. However, it is certainly unacceptable in many other cultures, including in the U.S. where

many tech companies are based [38]. Further, platforms should not take rankings of harmful content at face value, because a value being accepted does not make it right, and platforms should strive for more nuanced understandings beneath any stated value.

It is critical to note that neither the approach nor the findings in our study should be used as the sole reason to determine whether or not a piece of content should be moderated. Nor the findings alone determine the degree of harshness to which one should be punished for posting harmful content. we raise these precautions for two reasons: First, the question that why people perceived the harmful content as such remains unanswered. Second, more importantly, the perceptions of regular internet users should not be the gold standard of policymaking or the right way forward.

Implications for platform content moderation

The complexity of analyzing harmful content globally may be exacerbated by the inability to describe and interpret regionally, reflecting Fricker's [39] notion of hermeneutic injustice. Fricker defines hermeneutic injustice as when someone "has a significant area of their social experience obscured from understanding owing to prejudicial flaws in shared resources for social interpretation." In other words, a group of people may not have language or ability to describe and interpret a nuanced social experience that they do regularly experience. Fricker uses the example of stalking to demonstrate the concept of hermeneutic injustice. A man does not perceive stalking of a woman as harmful because the woman cannot describe her experience in the man's own interpretive system, and the man who never experienced stalking may have no idea what harm stalking actually entails. The same hermeneutic injustice is likely to apply in the context of content moderation as well. For example, people who are not aware of the consequences of anorexia nervosa may perceive its depiction as simply someone being "skinny" and therefore do not see it as being harmful.

Hermeneutic injustice also has broader consequences, because much of the platform policymaking relies on translation rather than local development. While platform policies do exist in multiple languages, as opposed to being developed organically by people across the world who are familiar with local contexts, they are often translated from English with a U.S. perspective, and essentially become different versions of U.S. ideals—not to mention how much information might be lost in the translation. The survey in our study suffers from the same epistemic limitation: It asked users to rate scenarios containing harmful content, but these scenarios, as well as the harmful content categories from which these scenarios arise, are often conceptualized and described by social media platforms in a small part of the world (in this case, the urban United States). People in the U.S. may perceive these categories and scenarios as comprehensive and representative, but it is possible that they do not even start to describe the harm and abuse that people in other parts of the world experience—it may not even be possible to describe them in English. It is also likely that people in other parts of the world cannot understand the harm that the U.S.-centric policies describe because they do not exist in their interpretive systems. Therefore, while our findings reveal that perceptions of harmful content can be different across countries, they should not be the predominant factor in policymaking. Content moderation practices cannot fully protect people if they are only imparting a specific idealized version of the harms these people experience. Policies for any country should be developed with local perspectives and in local terms.

Beyond hermeneutic injustice, sometimes users' values may simply go against the values that a platform or society holds, in which case it is in the platform's best interest to not take these user inputs into account in making platform policies. While research regarding social media platforms has historically valued soliciting user inputs and fulfilling them, there is no guarantee that user perceptions are always going to be beneficial (or even legal), especially in determining what content is welcome or unwelcome on a platform. More likely than not, platforms will not want to implement perspectives from users who want to see child pornography, who support ethnic cleansing, or who think anything that does not pertain to a certain religious belief should be removed. While these are extreme examples, some of the surprising findings in our study, such as the

depiction of "eating disorders" being lowly-rated in the United States, indicate that it is possible for some users to hold values that platforms may determine as harmful by value or by law. While these values may become less apparent when they are aggregated into large amounts of user inputs, but still have the potential to deprioritize harms that the platform may want to be prioritized.

Supporting information

S1 File. **Additional analysis details.**
(PDF)

S2 File. **Full survey instrument and participant response data.**
(ZIP)

Acknowledgments

The authors are grateful to Frank Kanayet, Charles Kiene, Joseph Seering, Kate Shores, and Amy Zhang for their help and feedback. Publication of this article was funded by the University of Colorado Boulder Libraries Open Access Fund.

Author Contributions

Conceptualization: Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, Jed R.Brubaker.
Data curation: Jialun Aaron Jiang.
Formal analysis: Jialun Aaron Jiang.
Funding acquisition: Casey Fiesler, Jed R. Brubaker.
Investigation: Jialun Aaron Jiang.
Methodology: Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, Jed R. Brubaker.
Project administration: Jialun Aaron Jiang.
Resources: Jialun Aaron Jiang.
Software: Jialun Aaron Jiang.
Supervision: Casey Fiesler, Jed R. Brubaker.
Validation: Jialun Aaron Jiang.
Visualization: Jialun Aaron Jiang.
Writing – original draft: Jialun Aaron Jiang.
Writing – review & editing: Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, Jed R. Brubaker.

References

1. Gillespie T. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media. Yale University Press; 2018.
2. Fick M, Dave P. Facebook's flood of languages leave it struggling to monitor content. Reuters. 2019;.
3. Facebook. Community Standards;. Available from: https://www.facebook.com/communitystandards/.
4. Smith HW. A modest test of cross-cultural differences in sexual modesty, embarrassment and self-disclosure. Qualitative Sociology. 1980;3(3):223–241. https://doi.org/10.1007/BF00987137

5. Roberts ST. Behind the Screen: Content Moderation in the Shadows of Social Media. Yale University Press; 2019.
6. Blackwell L, Ellison N, Elliott-Deflo N, Schwartz R. Harassment in Social Virtual Reality: Challenges for Platform Governance. Proc ACM Hum-Comput Interact. 2019;3(CSCW):Article 100. https://doi.org/10.1145/3359202
7. Chancellor S, Lin Z, Goodman EL, Zerwas S, De Choudhury M. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. CSCW'16. New York, NY, USA: ACM; 2016. p. 1171–1184. Available from: http://doi.acm.org/10.1145/2818048.2819973.
8. Gillespie T. Regulation of and by Platforms. In: The SAGE Handbook of Social Media. 55 City Road: SAGE Publications Ltd; 2018. p. 254–278. Available from: http://sk.sagepub.com/reference/the-sage-handbook-of-social-media/i2081.xml.
9. Klonick K. The New Governors: The People, Rules, and Processes Governing Online Speech. Harvard Law Review. 2018;131:1598–1670.
10. Chandrasekharan E, Pavalanathan U, Srinivasan A, Glynn A, Eisenstein J, Gilbert E. You Can'T Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. Proc ACM Hum-Comput Interact. 2017;1(CSCW):31:1–31:22. https://doi.org/10.1145/3134666
11. Jackson SJ, Gillespie T, Payette S. The Policy Knot: Re-integrating Policy, Practice and Design in Cscw Studies of Social Computing. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. CSCW'14. New York, NY, USA: ACM; 2014. p. 588–602. Available from: http://doi.acm.org/10.1145/2531602.2531674.
12. Fiesler C, Jiang JA, McCann J, Frye K, Brubaker JR. Reddit Rules! Characterizing an Ecosystem of Governance. In: Twelfth International AAAI Conference on Web and Social Media; 2018. Available from: https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17898.
13. Conger K. Twitter Will Ban All Political Ads, C.E.O. Jack Dorsey Says. The New York Times. 2019;.
14. Newton C. The secret lives of Facebook moderators in America; 2019. Available from: https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.
15. Chancellor S, Pater JA, Clear T, Gilbert E, De Choudhury M. #Thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. CSCW'16. New York, NY, USA: ACM; 2016. p. 1201–1213. Available from: http://doi.acm.org/10.1145/2818048.2819963.
16. Gerrard Y. Beyond the hashtag: Circumventing content moderation on social media. New Media & Society. 2018;20(12):4492–4511. https://doi.org/10.1177/1461444818776611
17. Seering J, Wang T, Yoon J, Kaufman G. Moderator engagement and community development in the age of algorithms. New Media & Society. 2019; p. 1461444818821316. https://doi.org/10.1177/1461444818821316
18. Fiesler C, Lampe C, Bruckman AS. Reality and Perception of Copyright Terms of Service for Online Content Creation. In: Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW); 2016. p. 1450–1461.

19. Satariano A. Facebook Can Be Forced to Delete Content Worldwide, E.U.'s Top Court Rules. The New York Times. 2019;.

20. Timmermann WK. Incitement in International Law. Routledge; 2014.

21. Scott M, Isaac M. Facebook Restores Iconic Vietnam War Photo It Censored for Nudity. The New York Times. 2016;.

22. Frey S, Krafft PM, Keegan BC. "This Place Does What It Was Built For": Designing Digital Institutions for Participatory Change. Proc ACM Hum-Comput Interact. 2019;3(CSCW):32:1–32:31. https://doi.org/10.1145/3359134

23. Barrett PM. Who Moderates the Social Media Giants? A Call to End Outsourcing. NYU Stern Center for Business and Human Rights; 2020.

24. Jiang JA, Middler S, Brubaker JR, Fiesler C. Characterizing Community Guidelines on Social Media Platforms. In: CSCW 2020 Companion; 2020.

25. Galanter E, Pliner P. Cross-Modality Matching of Money Against Other Continua. In: Moskowitz HR, Scharf B, Stevens JC, editors. Sensation and Measurement: Papers in Honor of S. S. Stevens. Dordrecht: Springer Netherlands; 1974. p. 65–76. Available from: https://doi.org/10.1007/978-94-010-2245-3_6.

26. Galanter E. The Direct Measurement of Utility and Subjective Probability. The American Journal of Psychology. 1962;75(2):208–220 https://doi.org/10.2307/1419604. PMID:13896303

27. von Hirsch A. Proportionality in the Philosophy of Punishment. Crime and Justice. 1992;16:55–98. https://doi.org/10.1086/449204

28. Fiedrich F, Gehbauer F, Rickers U. Optimized resource allocation for emergency response after earthquake disasters. Safety Science. 2000;35(1):41–57. https://doi.org/10.1016/S0925-7535(00)00021-7

29. Mitchell RC, Carson RT, Carson RT. Using Surveys to Value Public Goods: The Contingent Valuation Method. Resources for the Future; 1989.

30. Clement J. Facebook users by country 2019; 2019. Available from: https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/.

31. Cibelli K. The Effects of Respondent Commitment and Feedback on Response Quality in Online Surveys; 2017. Available from: https://deepblue.lib.umich.edu/handle/2027.42/136981.

32. Jolliffe IT. Principal Component Analysis. 2nd ed. Springer Series in Statistics. New York: Springer-Verlag; 2002. Available from: https://www.springer.com/gp/book/9780387954424.

33. Kriegel HP, Kröger P, Zimek A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Transactions on Knowledge Discovery from Data. 2009;3(1):1:1–1:58. https://doi.org/10.1145/1497577.1497578

34. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987;20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7

35. Schoenebeck S, Haimson OL, Nakamura L. Drawing from justice theories to support targets of online harassment. New Media & Society. 2020; p. 1461444820913122. https://doi.org/10.1177/1461444820913122

36. Irani L, Vertesi J, Dourish P, Philip K, Grinter RE. Postcolonial computing: a lens on design and development. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI'10. New York, NY, USA: Association for Computing Machinery; 2010. p. 1311–1320. Available from: https://doi.org/10.1145/1753326.1753522.

37. Peterson D, Panfil VR, editors. Handbook of LGBT Communities, Crime, and Justice. New York: Springer-Verlag; 2014. Available from: https://www.springer.com/gp/book/9781461491873.

38. Lawrence v. Texas, 539 U.S. 558 (2003); 2003. Available from: https://supreme.justia.com/cases/federal/us/539/558/.

39. Fricker M. Hermeneutical Injustice. Oxford University Press; 2007. Available from: https://oxford-universitypressscholarship-com.colorado.idm.oclc.org/view/10.1093/acprof:oso/9780198237907.001.0001/acprof-9780198237907-chapter-8.

3.8 Case Study 2

Offline events and online hate

Yonatan Lupu[1,2*], Richard Sear[3], Nicolas Velásquez[4], Rhys Leahy[1,2], Nicholas Johnson Restrepo[2], Beth Goldberg[5], Neil F. Johnson[2,3]

1 Political Science Department, George Washington University, Washington, DC, United States of America, 2 ClustrX LLC, Washington, DC, United States of America, 3 Physics Department, George Washington University, Washington, DC, United States of America, 4 Institute for Data, Democracy & Politics, George Washington University, Washington, DC, United States of America, 5 Google LLC, Menlo Park, California, United States of America

* ylupu@gwu.edu

Abstract

Online hate speech is a critical and worsening problem, with extremists using social media platforms to radicalize recruits and coordinate offline violent events. While much progress has been made in analyzing online hate speech, no study to date has classified multiple types of hate speech across both mainstream and fringe platforms. We conduct a supervised machine learning analysis of 7 types of online hate speech on 6 interconnected online platforms. We find that offline trigger events, such as protests and elections, are often followed by increases in types of online hate speech that bear seemingly little connection to the underlying event. This occurs on both mainstream and fringe platforms, despite moderation efforts, raising new research questions about the relationship between offline events and online speech, as well as implications for online content moderation.

Introduction

Online hate and extremism are critical problems across the globe. Online, groups of like-minded, prejudiced individuals create communities on social media platforms, where they recruit new members and sometimes coordinate offline behavior. Users' exposure to hate speech can have important negative consequences [1]. At times, this *online* activity can fuel high-profile *offline* violence, such as the January 6th attack on the U.S. Capitol. In turn, highly charged *offline* trigger events tend to be followed by sharp increases in *online* hate speech [2, 3]. Such spikes in *online* hate speech have, further, been shown to predict similar spikes in *offline* violent hate crimes [4–9].

Insights from literatures on social movements can be applied to online extremist communities. These movements typically focus on the creation of a group identity, particularly framed to the exclusion of others [10, 11]. Hate speech plays a particularly important role by intensifying feelings of in-group cohesion and negative feelings toward the out-group [12]. By communicating with each other on social media, individuals affirm their own identities as well as those of the group, often by using specific language, symbols, and images [13, 14]. They use social media to communicate, mobilize, and coordinate collective action [15, 16]. These technologies may facilitate such collective action and make violence more likely [17, 18]. Elites, including the media, play a particularly important role in the process. Like other individuals, members of online hate communities take cues from media and other elites regarding both which issues and events to focus on and the opinions they might hold on those issues and events [19–24]. By making some offline events more salient, elites can thus focus the attention of online hate communities, so the relationship between

https://doi.org/10.1371/journal.pone.0278511 Used with CC 4.0 license

offline trigger events and online hate speech is likely mediated in part by these actors. There is therefore a complex relationship between offline trigger events, elite/media cues and salience, online hate speech, and ensuing extremist violence that scholars across disciplines have sought to analyze and understand.

We study the connection between online hate speech and offline trigger events across the online ecology of moderated and unmoderated platforms. We focus on the following questions: How does online hate speech react to offline trigger events? Do these reactions differ across mainstream platforms that remove some hateful speech versus fringe platforms that are less likely to do so? Are offline events followed by online hate speech that is directly related to those events? Or, once hate speech has been kindled by an offline event, do members of online hate communities respond by using hate speech against seemingly unrelated targets?

Prior work has made significant progress on detecting and analyzing online hate speech. [Excellent surveys of this literature are available at [25, 26]]. With few exceptions [27, 28], existing work focuses on either classifying online hate speech as binary (i.e., hate or not hate) or classifying a small set of hate speech types [5, 29–31]. Most prior work tends to draw data from 1 or 2 social media platforms, usually mainstream and moderated platforms [31–36]. A recent survey found that Twitter was by far the most studied source in articles using automated detection of online hate speech [26]. Such approaches have two limitations we hope to mitigate. First, hate speech is likely to vary systematically between moderated and fringe platforms. Second, a more comprehensive classification of hate speech types is needed to better understand patterns of online hate speech following offline events.

We conduct a more comprehensive study of online hate speech by including a range of social media platforms and classifying a large set of hate speech types. We conduct a supervised machine learning analysis of 7 types of online hate speech using 59 million English-language posts drawn from 1,150 online hate communities on 6 social media platforms over the course of 19 months. This is a broad sample of online communities, particularly relative to prior work that tends to draw data from relatively few communities on moderated platforms. Our approach has several advantages. First, we include actively moderated mainstream platforms—Facebook, VKontakte, and Instagram—that have and enforce (to varying degrees) policies against hate speech, as well as the less-moderated platforms Gab, Telegram, and 4Chan. This allows us to capture online hate speech that would be filtered out of moderated platforms. Second, we collect data on posts on online hate communities across these platforms [37–39], allowing us to capture online hate speech where it is most likely to arise both before and after offline trigger events. Finally, by classifying a broad range of types of hate speech by their targets, we are able to analyze the extent to which offline events are followed by different types of online hate speech.

Data

Our data collection effort begins by identifying online hate communities, which are online forums in which hate speech is most likely to be used. Many social media platforms offer individuals the means to join communities of like-minded users, such as fan pages on Facebook, channels on Telegram, and anonymous message boards on 4Chan. We identified candidate online communities by using the sampling procedure described in greater detail in the SI. Our team manually searched social media platforms for hate communities and search their content to identify new communities. We included a community in our data as a hate community if 2 or more of the 20 most recent posts on the community included hate speech. For this step in the process, we defined hate speech as either (a) content that would fall under the provisions of the United States Code regarding hate crimes or hate speech according to Department of Justice guidelines; or (b) content that supports or promotes Fascist ideologies or regime types (i.e., extreme nationalism and/or racial identitarianism). The determination of whether an online community is a hate community was made manually using these criteria. We identified 1150 online hate communities that are included in this study.

https://doi.org/10.1371/journal.pone.0278511

Between June 1, 2019, and December 31, 2020, we captured the posts on these hate communities using a combination of automated and manual methods. We removed non-English-language posts by using Google's Compact Language Detector 2. Our data include the texts of 59 million posts, and do not include images, videos, and audio. The SI provides information on the distribution of posts and hate communities across the six platforms. We do not collect data on the geographic locations of users. A large share of them appears to be located in the United States based on their discussion of U.S. politics, but there are also frequent discussions of European issues and those of other English-speaking countries. We make no assumptions about the geographic locations of these users.

All data we collected were publicly posted on the Internet, and we had no contact with the users who posted the data. All data used in this study were collected and used in compliance with the terms of use of the online platforms. No personal identifying information was collected. No information about users was collected. User names were not collected. All the data are anonymous. We obtained advanced approval from the IRB.

Our study codes for 7 types of hate speech: speech targeting race, gender, religion, gender identity/sexual orientation (GI/SO), immigration, ethnicity/identitarian/nationalism (E/I/N), as well as anti-semitism. The SI provides additional information about how we define these types of hate speech. We chose these types of hate speech because, in our manual review of the hate communities, these types were both prevalent and distinguishable from each other (although they do co-occur).

Methods

To construct our training data, four annotators were trained to manually code whether or not a post contained each of the 7 types of hate speech. We trained the enumerators to distinguish between mere discussion of an issue and hate speech, e.g., to distinguish between non-hateful discourse about race and racial hate speech. A post could, in principle, be coded as containing between 0 and 7 types of hate speech. In total, the training data include 31,323 manually coded posts; the SI provides additional information about the coding procedure and training data.

Using the algorithm developed by [40], we split these data into 26,354 posts used to train the machine and 4,969 used to test performance. We compared the performance of 8 model architectures with differing classification and text representation methods. Using each model architecture, we trained 7 distinct models, one for each type of hate speech, each of which separately learns from the training data. Each of these 7 models provides a binary coding indicating whether or not a given post contained the applicable type of hate speech or not. We then compared the performance of these 7 models across the 8 model architectures in order to choose the best-performing architecture. The best performing model architecture used a neural network classifier with BERT embeddings [41]. The accuracy ranged from 91.7% to 98.3%, depending on the type of hate speech. We validated the machine results using human annotators and found that the results were highly reliable. The results below are based on the 7 models using this model architecture. The SI provides additional information regarding the 8 model architectures we tested, a comparison of their performance, a detailed description of the performance of the model architecture we chose, and detail regarding the validation exercise.

Results

We begin our analysis by focusing on the distribution of types of hate speech within the data. Panel A of Fig 3.7 shows how many posts contained each type of hate speech. By far the most common type of hate speech in the communities we tracked is racism, which appears in nearly 7 million posts. Most posts, despite originating from hate communities, did not contain any of the seven types of hate speech, as shown in Panel B. These communities often discuss unrelated issues, and some posts featured content other than text, which we did not analyze. About 20% of hate posts contained more than one type of hate speech. Panel C depicts how frequently each pair of hate speech types co-occurred within the same post. Most pairs are either

https://doi.org/10.1371/journal.pone.0278511

weakly positively or weakly negatively correlated. The largest positive correlation is between race and E/I/N (0.2396), while the largest negative correlation is between gender and anti-semitism (-0.0226).

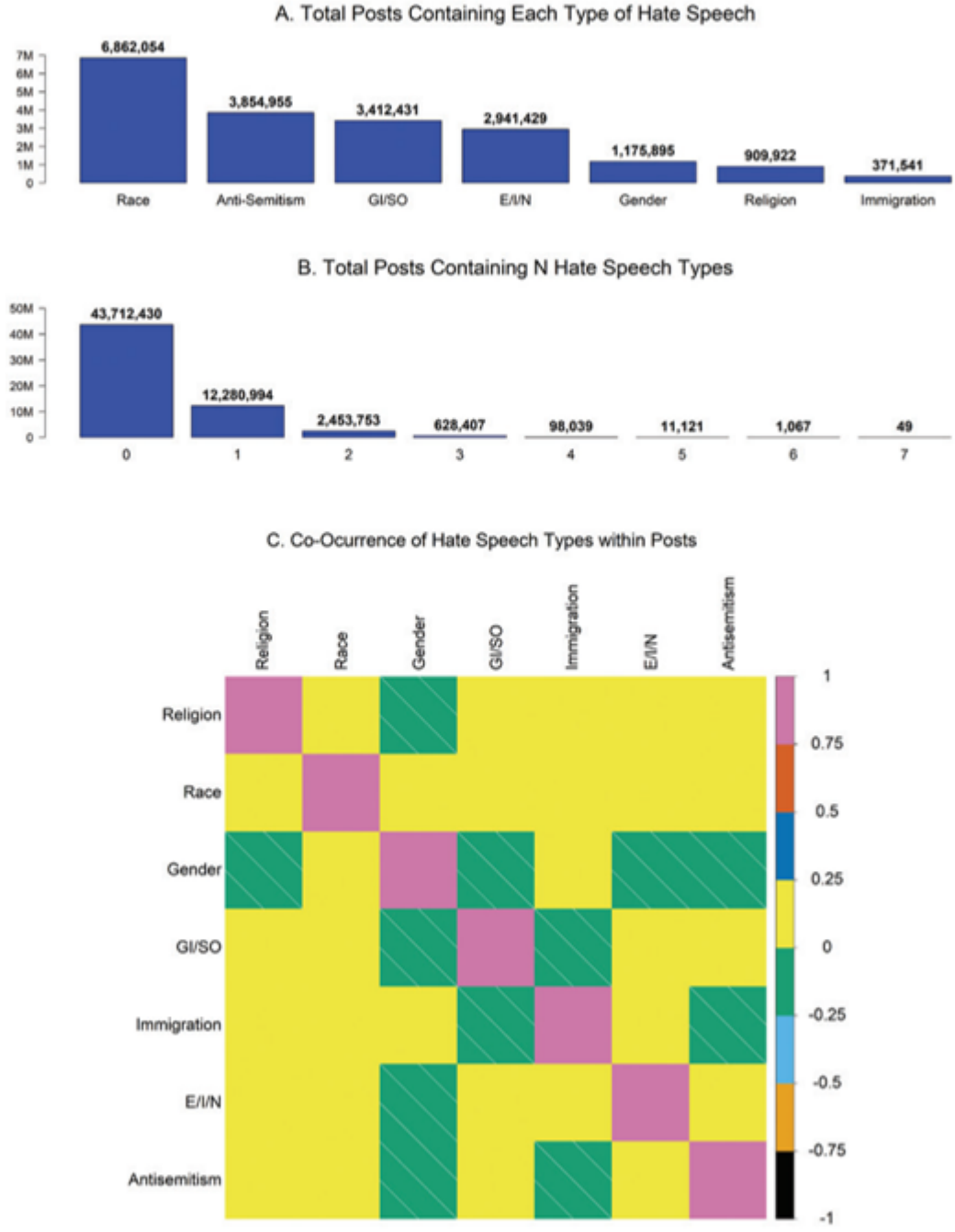


Fig 3.7 Posts, hate posts, and types of hate speech. Panel A shows how many posts contain each type of hate speech. Panel B shows how many posts contain different quantities of hate speech types. Panel C shows the pairwise correlations between hate speech types within posts. Diagonal lines indicate negative correlations.

https://doi.org/10.1371/journal.pone.0278511.g001

We next analyzed how hate speech changed over time. Fig 3.8 illustrates these changes in hate speech in these communities over time. The SI provides versions of these plots by platform. Panel A shows the temporal trend in the total count of daily posts across all 1150 communities. Activity in these communities differs systematically based on the day of the week, so here and elsewhere our results depict 7-day rolling averages. The total number of posts demonstrates an upward trend over the study period, increasing by about 67% from about 60,000 to 100,000 daily posts.

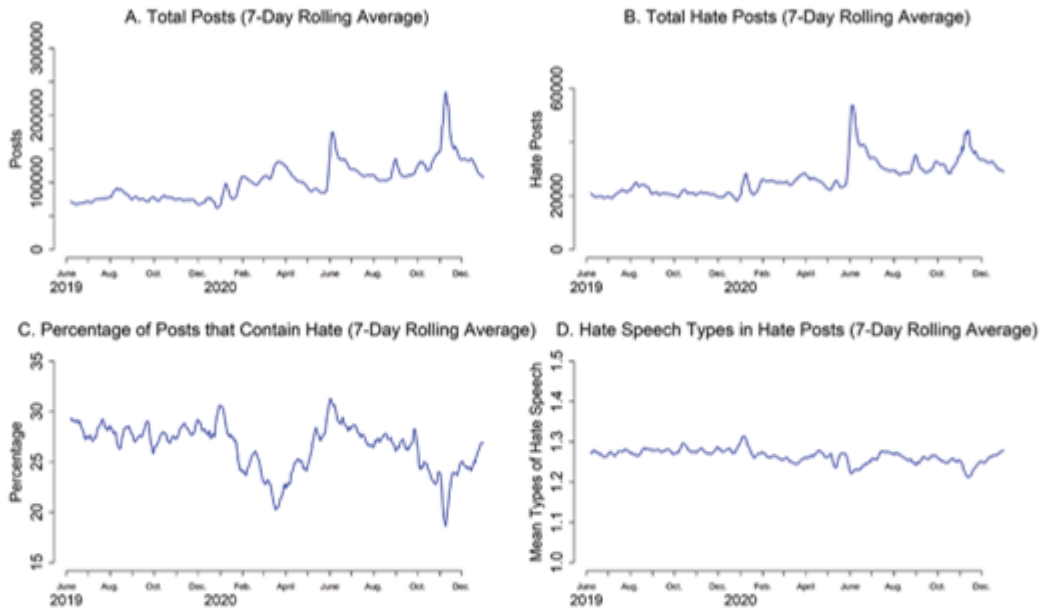


Fig 3.8 Temporal trends in posts and hate posts. Panel A shows the daily posts in the hate communities we tracked over time. Panel B shows the daily hate posts in these communities over time. Panel C shows the daily percentage of posts that were hate posts over time. Panel D shows the mean number of hate speech types in hate posts over time.

https://doi.org/10.1371/journal.pone.0278511.g002

There are two especially notable spikes in total activity: in late May/early June 2020, and in early November 2020. They key offline events discussed in the posts during these periods were (1) the May 25 killing of George Floyd by Minneapolis police officers and the ensuing protests organized by the Black Lives Matter (BLM) movement; and (2) the November 2020 U.S. elections. Panel B shows that not only were there far more posts than usual during these two events, there were also increases in hate posts. Most importantly, these results indicate that during periods of intense and contentious offline activity, the activity in online hate communities increases dramatically. In part this be due to the heightened salience of these events more generally, i.e., these events were likely being discussed not only in these hate communities, but elsewhere online, leading to greater online activity in general, including hate speech activity.

Panel C, however, indicates that these two events had differing effects on the relative frequency of hate speech within these communities. During the early days of the BLM protests, the percentage of posts that contained hate speech increase, meaning that hate posts increased not only in absolute terms (as shown in Panel B) but also in relative terms compared to other posts. During the period around the U.S. elections, however, Panel C shows that the percentage of posts that contained hate speech decreased. There was an increase in hate posts (Panel B), but in relative terms their frequency decreased, i.e., posts increased overall at a greater rate than hate posts. Finally, Panel D shows that the mean number of hate speech types contained in hate posts has remained quite consistent over time. Two relatively small decreases occur during the events mentioned above.

We continue analyzing how online hate speech changed over time by focusing on each type of hate speech separately. Fig 3.9 shows the temporal trends in each type of hate speech. The SI provides versions of these plots by platform. Many events, occurring both offline and online, can affect online hate speech; here, we focus on the largest changes in the time-series during the study period. The largest spikes in online hate activity occurred around the following dates in 2020: (1) January 3 with respect to religion and antisemitism; (2) February 29 with respect to immigration; (3) late May, with respect to race, gender, GI/SO, E/I/N, and antisemitism; and (4) early November, with respect to religion, gender, GI/SO, and immigration. Structural break analysis confirms that these changes in the data were not in line with prior patterns of hate speech ($p < 0.001$) and are the largest breaks in the time-series.

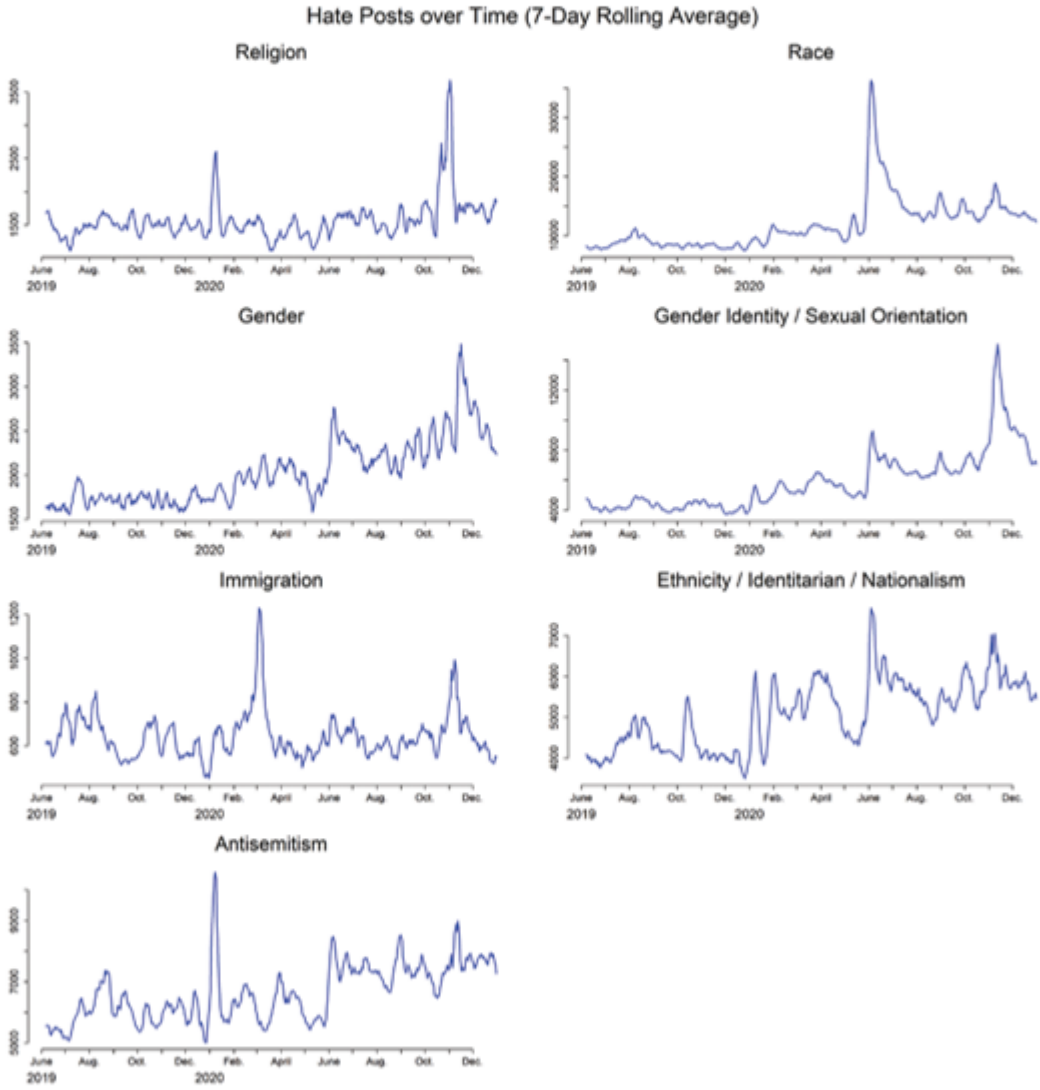


Fig 3.9 Temporal trends in each type of hate speech. Each panel shows the total daily posts that contain a given type of hate speech over time. Some types of hate speech appear much more often than others (Fig 3.9A), so the differences in y-axis scales (i.e., number of posts) should be noted.

https://doi.org/10.1371/journal.pone.0278511.g003

We manually reviewed random samples of 500 hate posts from each of the periods surrounding these spikes and concluded they are associated with the following offline events: (1) the assassination of Iranian General Qasem Soleimani; (2) a crisis at the border of Turkey and Greece that gained publicity when a large number of mostly Syrian refugees sought entry to the European Union; (3) the George Floyd murder and ensuing protests; and (4) the November 2020 U.S. elections. Events 1 and 2 were less highly publicized than the latter two, and their effects on our data may be less intuitive. We therefore supplemented our manual analysis with an automated topic analysis to confirm the spikes in hate posts were related to those events. See the SI for details.

We focus our next analysis on these events. Each panel in Fig 3.10 shows the changes in online hate speech following one of these events. Because different types of hate speech are used at different rates, we normalize the level of each type of hate speech at 100 based on the seven-day rolling average prior to the event. Fig 3.10 therefore shows the relative changes in the frequency of each type of hate speech after each event. Following the Soleimani assassination, several types of hate speech increased significantly, with the largest increases in religion (90%) and anti-semitism (65%). Interestingly, the religious hate expressed in these posts was largely Islamophobic, whereas the rise in anti-semitism focused on speculation that Israel was behind the assassination. The change in hate speech activity during the Turkey-Greece border crisis is the narrowest of the events we studied: a large increase in anti-immigration hate speech, with relatively little change in other types.

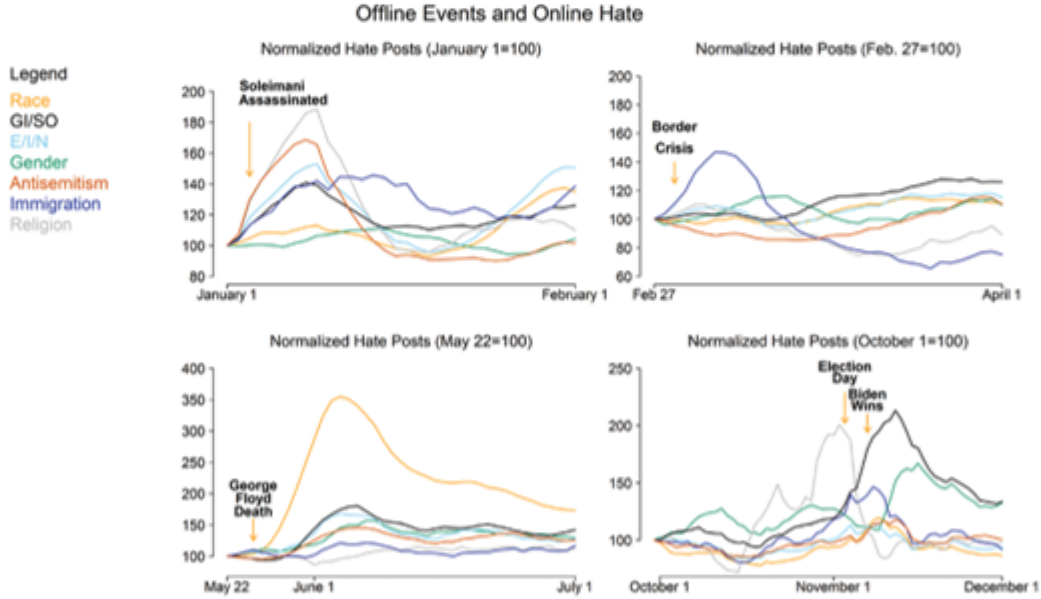


Fig 3.10 Offline events and online hate. Each panel shows the changes in the relative levels of each type of hate speech following a trigger event.

https://doi.org/10.1371/journal.pone.0278511.g004

The largest increases in hate activity during the study period followed the death of George Floyd and continued during the ensuing protests. The rate of racial hate speech increased by 250% by early June and, by the end of the year, remained about double what it had been prior to this event. This is perhaps not surprising given the negative reaction to the protests in some segments of the public. What is surprising, however, is that the frequency of most other types of hate speech also increased dramatically at this time, especially GI/SO (75%), E/I/N (60%), gender (50%), and anti-semitism (40%). While Floyd's death and the protests were largely framed as being about racial issues, it is not intuitively clear what relationship these events have to issues of, for example, gender and sexual orientation.

The 2020 U.S. elections also present an interesting case. Unlike the other events we have examined, the election was scheduled, so online communities might react in anticipation of it. Thus, in the fourth subfigure of Fig 3.10, we normalize hate speech levels a month prior to the election rather than a few days prior. Immediately before the election there was large increase in religion hate, most of which is Islamophobic; this appears to be in response to several events in Europe and not related to the election. By Election Day (November 3) and continuing through the declaration of Joe Biden as winner by the Associated Press (November 7), there were large increases in GI/SO hate speech (over 100%) and immigration hate speech (50%); a wave of gender hate speech soon followed. In some ways, these waves of hate speech are surprising. Issues of gender identity and sexual orientation were not prominent during the campaign; the ensuing wave of associated online hate speech appears to be largely a case of users employing anti-LGBTQ slurs in a generalized manner to malign a wide range of political targets, such as candidates, parties, and voters. Likewise, immigration was a less prominent campaign issue than it had been in 2016, but after Biden's victory online hate speech increased due to speculation that he would moderate U.S. immigration policy. The ensuing wave of gender hate speech appears to be at least partly targeted toward Vice President Kamala Harris based on our manual review of the relevant posts. Interestingly, although racial issues were at the forefront during the campaign, online race hate speech changed relatively little before and after the election. This may be because the rate of race hate speech, despite having dropped from its high in June, remained elevated throughout this period as compared to the period prior to George Floyd's death (see Fig 3.9).

Because hate activity increased to the greatest extent after Floyd's death, we focus in more detail on this event. Fig 3.11 shows results similar to those of Fig 3.10, but broken down by platform. Instagram is excluded from Fig 3.11 because there were not enough daily hate speech posts during this period to make meaningful inferences about changes in activity. The pattern on 4Chan largely follows the overall pattern across all the platforms because 4Chan makes up the majority of our data. More surprising is the relatively large increase in hate speech on Facebook. On Facebook, the daily average of posts containing race hate speech increased by a factor of 10 within days of Floyd's death and remained highly elevated (5x) a month later. Other types of hate speech, especially anti-semitism, gender, and GI/SO, also increased by large amounts on Facebook during these offline events. We observe similar patterns on Gab and Telegram: a very large increase in race hate speech along with smaller, but nonetheless notable, increases in several other types of hate that endure several weeks after the event. On VKontakte, however, there was relatively little reaction to these events, perhaps because users there are less attuned to U.S. politics. The increase in GI/SO hate activity on VKontakte in mid-June appears to be unrelated to these events.

https://doi.org/10.1371/journal.pone.0278511

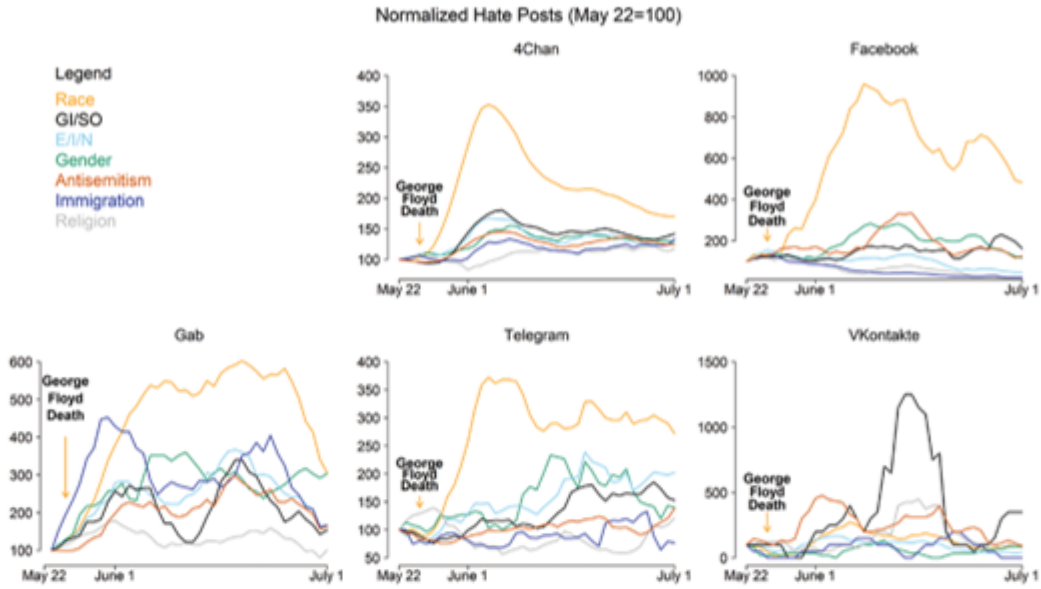


Fig 3.11 Hate speech by platform after George Floyd's death. Each panel shows the changes in the levels of each type of hate speech following the death of George Floyd on a given platform.

https://doi.org/10.1371/journal.pone.0278511.g005

Discussion

Our findings have several implications for researchers and practitioners. The first is that online hate speech is a large and growing problem that affects both mainstream and fringe platforms. Over the study period, the daily volume of hate speech posts began at an average of 20,000, peaked at around 50,000, and ended at around 25,000. Some prior work has found that the volume of extremist content and activity declined on YouTube, a moderated platform [42, 43]. Yet we find that, on the platforms we studied, including those that are moderated, hate speech increased despite elevated efforts to curtail it by moderated platforms, particularly during the second half of 2020. Although some platforms are increasingly committed to removing hate speech and other malicious content, hate content persists on such platforms (see, for example, Fig 9 in S1 File). Despite efforts by mainstream platforms to remove hate speech, it has grown in volume on fringe platforms such as 4Chan and Telegram (see Figs 8 and 10 in S1 File). Online extremists are known to migrate to less-moderated platforms after being removed from mainstream platforms [38, 44], which may be less susceptible to public and policymaker pressure.

It is therefore crucial to continue to analyze how mainstream platforms' content moderation affects the overall ecology of online hate speech. In particular, our findings raise several research questions. First, given the prior findings of decreases of extremist content on YouTube and our finding of increases in hate speech on other moderated platforms, we need a better of understanding of the conditions under which platforms can be more effective in moderating content. Key factors appear to include the type of content (e.g., hate speech as compared to other extremist content), the format of the content (e.g., text versus video), and the varying responsiveness of malicious content to offline events. In addition, given how quickly users are known to migrate across platforms, a key issue for future research is the overall impact of this migration on malicious content. When extremists and others who post malicious content are deplatformed from mainstream platforms, how many of them migrate to fringe platforms? Do only the most invested and extreme users do so, or do they bring a more casual following along with them?

After such migration, does the nature of malicious content, such as hate speech, change? Does this content become more extreme, hateful, or otherwise malicious, or not?

Secondly, the spikes in online hate speech following offline events are not limited to fringe platforms. We know from prior work that online hate speech can react strongly following offline events [2], but our data allow us to compare the sizes of these increases across platforms and types of hate speech, which yields important new results. After the death of George Floyd, for example, the largest increase in online hate speech in absolute terms (i.e., the number of hate speech posts) was on 4Chan, but in *percentage terms* the largest increase was on Facebook (see Fig 3.11). This is a particularly striking result given Facebook's increased attention to content moderation. The implication of this for moderated platforms is that they should be especially watchful during such periods, but also that increases in hate speech on other, fringe platforms may be predictive of potentially similar increases on their own platforms, a point we hope to explore in future research.

Finally, and perhaps most importantly, we find that online hate speech reacts to differing offline events in different and sometimes counter-intuitive ways. We might conceptualize at least three types of offline events in terms of the patterns of hate speech that follow them. First would be offline events that could plausibly be followed by significant increases in hate speech (i.e., events involving frequent targets of hate speech), but that draw little reaction from online hate communities. For example, on March 13, 2020, Louisville, Kentucky, police fatally shot an African-American woman named Breonna Taylor in her home. The killing was followed by a relatively small discussion in the online hate communities we tracked–in contrast to the large reaction to the murder of George Floyd two months later. A second type of offline event is one that is followed by a discrete increase in online hate speech that appears to be directly related to the offline event, and then dissipates relatively quickly. The assassination of Suleimani and the Turkish-Greek border crisis appear to be events of this type. A third type of offline event is followed by sharp increases in and broad cascades of online hate speech that build on, and potentially transform, the meaning and interpretation of the initial offline event and target groups in an extensive and seemingly indiscriminate manner. Events of this type can have longer-lasting effects on patterns and volume of online hate speech. In our study period, the murder of George Floyd and the U.S. Elections appear to be events of this type. It may be intuitive, for example, that online hate speech regarding race would increase during the BLM protests given the way those events were centered on the subject of race. Yet most other types of online hate speech also increased during that time, including, for example, GI/SO hate, which, on the surface, would appear to have an indirect connection to issues of police brutality and race.

Conclusions

What factors explain the different relationships between individual offline events and online hate speech? A detailed analysis of this question is beyond the scope of this paper, but media and elite attention likely play an important role, as discussed in the introduction. Both the volume and variety of online reactions to offline events depend, in part, on the salience of those events in other media. For example, the killing of Taylor received little attention in the U.S. national media in the weeks that followed it (perhaps because it occurred during the time when many U.S. jurisdictions were announcing the first lockdowns of the COVID-19 pandemic), and became significantly more salient during the protests that followed the Floyd murder. The Soleimani assassination and EU border crisis received media attention, but for a shorter time span than the BLM protests and U.S. election. These observations suggest that media attention may be one of the key factors that explains why the patterns of online hate speech following individual offline events differ so sharply.

https://doi.org/10.1371/journal.pone.0278511

Yet this factor likely does not explain the range of our findings. For example, if online racist hate speech was, in part, activated by media attention to the BLM protests that made racial identity a more salient issue, why did other forms of hate speech also increase dramatically at that time? Why does this broad activation of online hate speech occur following some offline events but not others? We hope future work will investigate these questions. A particularly important question for future work is to investigate which *types of offline events* are more likely to be followed by broad and seemingly indiscriminate cascades of online hate speech. Likewise, our results also raise the question of which *types of online hate speech*, once triggered by such events, are more likely to be followed by such a cascade into other types of hate speech.

Our study faces two important limitations. It is not possible to identify the entire population of online hate communities and sample from it. The online ecology of social media platforms is vast and constantly changing. Tracking and capturing data on these platforms is also limited by terms-of-service and privacy concerns. It is therefore not possible to capture the entire ecology of online speech nor to confidently claim that any sample of such speech is representative of the ecology. Our list of online hate communities is nonetheless the most comprehensive such list we are aware of. After several months of search by several members of our team, we are not aware of any other communities on the 6 platforms we studied that meet our definition of hate communities. We have captured many of the most prominent and active online communities on which hate speech is regularly used. Our findings and conclusions are limited to the public communities we identified. We can infer from our results the relationship between offline events and the use of hate speech on our sample of prominent online hate communities, but we cannot infer from our results, nor from those of prior work, whether or not this relationship holds across the online ecology.

A second limitation of our work is that our analysis does not allow for causal inference between observed offline and online phenomena. We reviewed the posts associated with the spikes in online hate speech discussed above and concluded they are substantively related to the trigger events we described. Following the death of George Floyd protests, for example, the large increase in posts that contained race hate speech (see Fig 3.10) included many posts that referenced Floyd, the circumstances surrounding his death, and the ensuing protests, leading us to conclude the increase in online hate speech was related to these offline events.

We conclude by noting that much work remains to be done to better understand the relationship between offline events and online hate speech. We have conducted a wide-ranging analysis of online hate speech–across types of hate speech, many online communities, and a diverse set of platforms. We have provided evidence that suggests complex relationships between offline events and online hate, between different types of online hate, and between moderated and fringe platforms. These results raise important questions for scholars across multiple disciplines and suggest several strategies for mitigating online hate speech. First, content moderators should be increasingly watchful of offline events, and efforts may need to be augmented following highly salient events. Second, the effects of an offline event on online hate speech may be difficult to predict, and may quickly evolve and broaden, which further suggests a broad and rapid mitigation effort in response. Because online hate speech levels are predictive of offline hate crimes [4–9], the potential exists for an offline trigger event to be followed by a broad cascade of online hate speech that, in turn, precedes hate crimes of a nature unrelated to the initial event. Third, because the effects of an offline event on different types of platforms are often similar, content moderators on mainstream platform could advance their efforts by systematically observing ongoing developments on other, fringe platforms.

Supporting information

S1 File. **Supplementary material to the manuscript.**
(PDF)

https://doi.org/10.1371/journal.pone.0278511

Author Contributions

Conceptualization: Yonatan Lupu, Nicolas Velásquez, Beth Goldberg.
Data curation: Richard Sear, Rhys Leahy, Nicholas Johnson Restrepo.
Formal analysis: Yonatan Lupu.
Funding acquisition: Yonatan Lupu, Neil F. Johnson.
Investigation: Yonatan Lupu, Nicholas Johnson Restrepo.
Methodology: Yonatan Lupu, Richard Sear, Nicolas Velásquez.
Project administration: Yonatan Lupu, Beth Goldberg.
Resources: Yonatan Lupu.
Software: Richard Sear, Nicolas Velásquez.
Supervision: Yonatan Lupu, Beth Goldberg, Neil F. Johnson.
Validation: Richard Sear, Rhys Leahy.
Visualization: Yonatan Lupu.
Writing – original draft: Yonatan Lupu, Rhys Leahy.
Writing – review & editing: Yonatan Lupu.

References

1. Oksanen A., Hawdon J., Holkeri E., Näsi M., & Räsänen P. (2014). Exposure to online hate among young social media users. In *Soul of society*: *a focus on the lives of children & youth*. Emerald Group Publishing Limited.
2. Burnap P., & Williams M. L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data science*, 5, 1–15. https://doi.org/10.1140/epjds/s13688-016-0072-6 PMID:32355598
3. Siegel, A., Tucker, J., Nagler, J., & Bonneau, R. (2018). Socially mediated sectarianism. *Unpublished manuscript*. https://alexandra-siegel.com/wp-content/uploads/2019/05/Siegel_Sectarianism_January2017.pdf
4. Müller, K., & Schwarz, C. (2020). From hashtag to hate crime: Twitter and anti-minority sentiment. *SSRN 3149103*.
5. Williams M. L., Burnap P., Javed A., Liu H., & Ozalp S. (2020). Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 60(1), 93–117.
6. Benesch S., Buerger C., Glavinic T., Manion S., & Bateyko D. (2018). Dangerous speech: A practical guide. *Dangerous Speech Project*.
7. Müller K., & Schwarz C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4), 2131–2167.
8. Karell D., Linke A., and Holland E. C. 2021. "Right-Wing Social Media and Political Unrest," SocArXiv.
9. Olteanu, A., Castillo, C., Boy, J., & Varshney, K. (2018, June). The effect of extremist violence on hateful speech online. In *Proceedings of the international AAAI conference on web and social media* (Vol. 12, No. 1).
10. Hunt S. A., & Benford R. D. (1994). Identity talk in the peace and justice movement. *Journal of Contemporary Ethnography*, 22(4), 488–517.

11. Snow, D. (2001). Collective identity and expressive forms. https://escholarship.org/uc/item/2zn1t7bj;jsessionid&.
12. Vollhardt J., Coutin M., Staub E., Weiss G., & Deflander J. (2006). Deconstructing hate speech in the DRC: A psychological media sensitization campaign. *J. Hate Stud.*, 5, 15.
13. Hine, G. E., Onaolapo, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Samaras, R., et al. (2017, May). Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Eleventh International AAAI Conference on Web and Social Media*.
14. Klein O., Spears R., & Reicher S. (2007). Social identity performance: extending the strategic side of SIDE. *Personality and Social Psychology Review*, 11(1), 28–45. https://doi.org/10.1177/1088868306294588 PMID:18453454
15. Ackland R., & O'Neil M. (2011). Online collective identity: the case of the environmental movement. *Social Networks*, 33(3), 177–190.
16. Polletta F., & Jasper J. M. (2001). Collective identity and social movements. *Annual Review of Sociology*, 27(1), 283–305.
17. Pierskalla J. H., & Hollenbach F. M. (2013). Technology and collective action: The effect of cell phone coverage on political violence in Africa. *American Political Science Review*, 107(2), 207–224.
18. Weidmann N. B. (2015). Communication networks and the transnational spread of ethnic conflict. *Journal of Peace Research*, 52(3), 285–296.
19. Zaller J. R. (1992). *The nature and origins of mass opinion*. Cambridge university press.
20. Druckman James N. 2001. "The Implications of Framing Effects for Citizen Competence." Political Behavior 23 (3): 225–56
21. Siegel A. A., & Badaan V. (2020). #No2Sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review*, 114(3), 837–855.
22. Del Vicario M., Vivaldo G., Bessi A., Zollo F., Scala A., Caldarelli G., et al. 2016. Echo chambers: Emotional Contagion and Group Polarization on Facebook. *Scientific Reports*, 6(1), 37825. https://doi.org/10.1038/srep37825 PMID:27905402
23. McCombs M. E., & Shaw D. L. (1972). The agenda-setting function of mass media. *Public opinion quarterly*, 36(2), 176–187.
24. Salminen J., Sengün S., Corporan J., Jung S. G., & Jansen B. J. (2020). Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PloS one*, 15(2), e0228723. https://doi.org/10.1371/journal.pone.0228723 PMID:32084164
25. Schmidt, A., & Wiegand, M. (2017, April). A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media* (pp. 1–10).
26. Fortuna P., & Nunes S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4), 1–30.
27. Silva, Leandro, Mainack Mondal, Denzil Correa, Fabricio Benevenuto, and Ingmar Weber. 2016. "Analyzing the Targets of Hate in Online Social Media." *ArXiv:1603.07709 [Cs]*, March. http://arxiv.org/abs/1603.07709.

28. Alorainy, Wafa, Pete Burnap, Han Liu, and Matthew Williams. 2018. "The Enemy Among Us: Detecting Hate Speech with Threats Based 'Othering' Language Embeddings." ArXiv:1801.07495 [Cs], March. http://arxiv.org/abs/1801.07495.

29. Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In Proceedings of the 24th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, pages 29–30.

30. Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, pages 145–153.

31. Burnap P., & Williams M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2), 223–242.

32. Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In Proceedings of the 21st ACM international conference on Information and knowledge management, pages 1980–1984, Maui, HI, USA. ACM.

33. Zhang Z., & Luo L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5), 925–945.

34. Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)* (pp. 86–95).

35. Berglind, T., Pelzer, B., & Kaati, L. (2019, August). Levels of hate in online environments. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 842–847).

36. Siegel A. A., Nikitin E., Barberá P., Sterling J., Pullen B., Bonneau R., et al. (2021). Trumping hate on Twitter? Online hate speech in the 2016 US election campaign and its aftermath. *Quarterly Journal of Political Science*, 16(1), 71–104.

37. Johnson N. F., Zheng M., Vorobyeva Y., Gabriel A., Qi H., Velásquez N., et al. (2016). New online ecology of adversarial aggregates: ISIS and beyond. *Science*, 352(6292), 1459–1463. https://doi.org/10.1126/science.aaf0675 PMID:27313046

38. Johnson N. F., Leahy R., Restrepo N. J., Velasquez N., Zheng M., Manrique P., Devkota P. & Wuchty S. (2019). Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, 573(7773), 261–265. https://doi.org/10.1038/s41586-019-1494-7 PMID:31435010

39. Johnson N. F., Velásquez N., Restrepo N. J., Leahy R., Gabriel N., El Oud S., et al. (2020). The online competition between pro-and anti-vaccination views. *Nature*, 582(7811), 230–233. https://doi.org/10.1038/s41586-020-2281-1 PMID:32499650

40. Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011, September). On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 145–158). Springer, Berlin, Heidelberg.

41. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

42. Munger K., & Phillips J. (2020). Right-Wing YouTube: A supply and demand perspective. *The International Journal of Press/Politics*, 1940161220964767.

43. Hosseinmardi H., Ghasemian A., Clauset A., Mobius M., Rothschild D. M., & Watts D. J. (2021). Examining the consumption of radical content on YouTube. *Proceedings of the National Academy of Sciences*, 118(32). https://doi.org/10.1073/pnas.2101967118 PMID:34341121

44. Mitts T. (2021). Countering Violent Extremism and Radical Rhetoric. *International Organization*, 1–22.

3.9 Case Study 3

RESEARCH ARTICLE

Examining the role of moral, emotional, behavioural, and personality factors in predicting online shaming

Shannon Raine Muir*, Lynne Diane Roberts, Lorraine Sheridan, Amy Ruth Coleman

School of Population Health, Curtin University, Bentley, Western Australia, Australia

* shannon.muir@curtin.edu.au

Abstract

Online shaming, where people engage in social policing by shaming perceived transgressions via the internet, is a widespread global phenomenon. Despite its negative consequences, scarce research has been conducted and existing knowledge is largely anecdotal. Using a correlational online survey, this mixed-method study firstly assessed whether moral grandstanding, moral disengagement, emotional reactivity, empathy, social vigilantism, online disinhibition, machiavellianism, narcissism, and psychopathy predict participants' (N = 411; aged 15–78) likelihood to engage in online shaming. Two hierarchical multiple regression analyses revealed these predictors significantly accounted for 39% of variance in online shaming intentions, and 20% of variance in perceived deservedness of online shaming (f^2 = .25 and .64 respectively, $p < .001$). A content analysis of an open-ended question offered further insights into public opinions about online shaming. These qualitative findings included the perception of online shaming as a form of accountability, the perceived destructive effects of online shaming, the perceived role of anonymity in online shaming, online shaming as a form of entertainment, online shaming involving 'two sides to every story', the notion that 'hurt people hurt people', online shaming as now a social norm, and the distinction between the online shaming of public and private figures. These findings can be used to inform the general public and advise appropriate responses from service providers and policy makers to mitigate damaging impacts of this phenomenon.

Introduction

The use of shaming to mitigate undesirable behaviours online has been evident since the internet was first popularised [1]. Whilst there is no widely accepted definition of online shaming, it can broadly be considered a phenomenon where people engage in social policing by shaming perceived transgressions via social media and other internet technologies [2]. Online shaming can take many forms, with common examples including people ridiculing discriminatory posts on social media, or recording and sharing photos or videos of people breaking social norms (e.g., unsafe driving) in public [1]. Much like other forms of shaming, at the heart of this vigilante behaviour lies expressions of social disapproval, with the outcome usually being remorse from the 'offender' and disapproval from peers [3].

Competing interests: The authors have declared that no competing interests exist.

While online shaming is largely depicted as having palpable and far-reaching negative impacts [4, 5], the perceived benefits to exposing people or organisations online have also been highlighted. For instance, online shaming affords everyday people a platform where they can voice their disapproval of perceived injustices and a multitude of inappropriate behaviours, where these behaviours may have previously remained unpunished

https://doi.org/10.1371/journal.pone.0279750 Used with CC 4.0 license

or unacknowledged [4]. Online shaming is said to empower individuals to enforce laws and prosocial behaviours, as well as contribute to overall societal reform, and greater institutional transparency [4, 5].

Online shaming typically involves exposing and distributing the content being shamed, followed by various displays of aggression (e.g., insults, threats) and outbursts [5]. Revealing the shamed individuals' personally identifiable information as a means of condemnation and punishment is also common, with online aggression sometimes evolving into offline harassment (e.g., stalking, vandalising property) afterwards [2, 4]. Whilst one person may have little influence, the collective power from many individuals combined can have a devastating and pervasive impact through the rapid and broad transmission of information online, essentially becoming an online mob trial without due process and no locational or geographical restrictions [2, 4].

Ingraham and Reeves [6, p.456] reflect on the destructive impact modern shaming can have, stating, "the police and the courts are often unable to mete out punishments as severe or intimidating as the ostracism, job loss, death threats, and physical attacks that can accompany what Urry (1999) calls our increasingly mediated culture of shame." Others discuss the isolation that surrounds the shamed individual and the "impulse to cover up and hide" [4, p.95]. There often is no chance for rehabilitation for those who are shamed, especially given the relative permanency and uncertain timeline of punishments in the online world, however those responsible for the mass shaming are said to often not consider or care about this [7]. More broadly, there is also the possibility that online shaming will contribute to generating an oppressive atmosphere in the online world. This may be amplified due to the absence of proper due processes and lack of guaranteed proportionality, potentially resulting in online shaming becoming increasingly likened to vengeance and bullying [4]. At present, research in this area is still in its infancy, with existing discussion regarding online shaming largely driven by the media and journalists rather than evidenced through empirical inquiry.

Shaming reimagined

The practice of public shaming has an extensive history, with notable examples of state enforced punishments including public whippings, branding, scarlet letters, and the stockade [4, 8]. Although these practices often served as an effective punishment and deterrent for the wider community, this was all but abolished in the nineteenth century due to being considered too cruel, humiliating, and largely inhumane [8]. Public shaming is said to remove the offender's dignity, and renders them unable to redeem themselves or ever re-enter 'normal' society [8, 9].

As for the online world, the proliferation of internet technologies in the twenty-first century (particularly the rise of social networking websites like Facebook, Instagram, and Twitter) has galvanised individuals' ability to capture and share their day-to-day lives instantly and with ease [10]. This has led to many members of the public engaging in social or peer surveillance as a form of social control [1], with involvement from authority figures, state officials, and mass media no longer a prerequisite of social norm [11]. Instead, individuals today have an unparalleled agency to observe, judge, and reprimand other internet users for perceived moral infractions [6]. This mass online surveillance has had a pivotal role in the resurrection of shaming as a form of punishment in contemporary society. Further, rather than being simply a re-emergence of public shaming on a contemporary medium, the internet has altered the prerequisites and social conditions needed for public shaming to occur. That is, it is no longer necessary for an actual crime to be committed, but now perceived deviations from social norms are enough to warrant public ridicule for the world to see [11].

Explaining online shaming

It has been argued that online shaming can be a way for individuals to indicate their non-support for certain socially offensive behaviours, effectively seeking social status by signalling to others their own credibility [12]. Labelled virtue signalling in mass media, and termed moral grandstanding in the literature, this factor

https://doi.org/10.1371/journal.pone.0279750

has been put forward as a theoretical explanation for online shaming [e.g., 13], but has yet to be examined empirically. Similarly, moral disengagement (convincing oneself that certain ethical standards do not apply to them due to perceived extenuating circumstances) has been demonstrated to be a consistent positive predictor of other acts of online aggression (e.g., cyberbullying) previously [14], but currently has not been tested in relation to online shaming specifically. Again based on comparable online behaviours and media reports of online shaming [15], other potentially useful but currently unexamined predictors of online shaming engagement include higher emotional reactivity [16] and lower empathy [17].

Research by Skoric et al [1] suggests people who shame others online often do not perceive themselves to be shaming individuals, but rather view their activities as a way of contributing to society by reinforcing social norms and deterring others from committing the same 'offence'. Online shaming is also seen as a way of exerting social control or 'social policing', with the aim of deterring deviance and fostering group solidarity [15]. It is the reinforcement of social norms and social control that differentiates online shaming from other activities such as cyber-harassment or online bullying [18]. The intention behind online shaming is said to be to identify and punish social norm violations [1], whereas cyberbullying is instead typically considered to be a repeated personal online attack on an individual (making these online behaviours comparable, but ultimately distinct from each other). Online shaming can also be considered a type of social vigilantism, in that individuals are autonomous, acting spontaneously, and free from organisational or state input to restore justice on their own accord [18]. In this sense, online shaming can be seen as a type of positive punishment, with shaming enforced by everyday people, and transpiring as a reaction to individuals' dissatisfaction with already established methods of punishment (or lack thereof) for certain acts or behaviours [1].

Another proposed contributing factor of online shaming is online disinhibition [15], which is a phenomenon whereby people feel less restrained and are more willing to express themselves freely on the internet (e.g., harsher criticisms, threats) than in person [19]. A multitude of studies have empirically demonstrated increased aggression and hostile behaviours in computer-mediated interactions compared to face-to-face interactions [e.g., 20], which may also contribute to the 'mob mentality' often seen online. Similarly, various personality factors (e.g., psychopathy, sadistic traits) have been shown to predict the likelihood of engaging in aggressive online behaviours [21], and likewise may play a role in online shaming. Collectively, it is evident there are numerous possibly fruitful but currently untested avenues of investigation for researchers wishing to explain and predict online shaming.

Significance

To date, little evidence exists regarding the motivations for online shaming, which is surprising given the effects can be so devasting and uncontrollable. As previously mentioned, a lot of existing studies pose various concepts as explanations for online shaming without actually testing these, and current discussion on the impacts of online shaming is largely anecdotal or media-driven. The online world is fuelled by social norms, however the means through which these norms are regulated evolve rapidly and unpredictably, rendering online shaming an increasingly pervasive and relevant societal problem. Despite this, current legislative attempts to mitigate the negative effects of online shaming have faced harsh criticism. For example, whilst the European Union's 'right to be forgotten' article, which was put forth in the 'General Data Protection Regulation' [22], was met with some support, it largely faced disapproval and outrage [23]. Depending on the jurisdiction and the specific case, there may be several viable legal claims for the person being shamed online (e.g., defamation, criminal hate speech, and use of a telecommunications device to abuse, threaten or harass), however the person conducting the shaming is often not identifiable or there may be too many culpable parties to enact legal consequences against [5]. Whilst there have been cases where legal solutions have been afforded, such as successful defamation lawsuits [5], it is still largely unclear what legal actions are available to those who have been subjected to online shaming. It is essential to systematically analyse and disseminate

https://doi.org/10.1371/journal.pone.0279750

the motivators and underlying trends of online shaming to establish a more comprehensive understanding of this issue. By doing so, the findings can be used by policy makers to inform appropriate legislation, by service providers to advise appropriate responses, and by psychologists and other health professionals when working with victims of online shaming. Further, findings can also be dispersed to the public to encourage discourse based on empirical support rather than just theoretical and anecdotal speculation.

The present study

The overarching aim of the present study was to expand upon currently scarce understandings of, and provide empirical evidence for, the motivators behind participation in online shaming. In this paper, we firstly present the utility of several moral, emotional, behavioural, and personality factors in predicting online shaming engagement, followed by an exploratory qualitative account of public opinions regarding online shaming. For the quantitative aspect of this research, the following hypothesis was posited: moral grandstanding, moral disengagement, emotional reactivity, social vigilantism, online disinhibition, machiavellianism, narcissism, and psychopathy will be positive predictors, and empathy be a negative predictor, of likelihood to engage in online shaming.

Method

Research design

A cross-sectional, correlational design was employed using an online survey (with a small qualitative component embedded via an open ended-question). The quantitative predictors were moral grandstanding, moral disengagement, emotional reactivity, empathy, social vigilantism, online disinhibition, machiavellianism, narcissism, and psychopathy. The criterion was the likelihood to engage in online shaming. Social desirability was also included as a control variable to statistically account for participants providing responses they may deem to be more desirable. The purpose of the qualitative component was to afford participants the opportunity to share additional thoughts about online shaming not captured within the quantitative strand. The qualitative and quantitative findings have been integrated in the discussion section where appropriate, via a weaving narrative approach.

Participants

A convenience sample of 411 participants (311 women, 96 men, and 4 identifying as another gender) between 15 and 78 years old ($M = 24.76$, $SD = 9.59$) completed this research, with an average of 14.73 ($SD = 3.30$) years of formal education, and almost half (49%) identifying as Caucasian. The majority of respondents were from Australia (77%), were students (75%), and were employed (60%). The most commonly reported religions were none or Atheist (44%), Catholic (15%), and Christian (15%). On average, participants spent approximately six hours online daily ($SD = 2.86$) and approximately three hours on social media daily ($SD = 2.14$). About 20% ($n = 84$) of participants indicated they had been shamed online before, around 62% ($n = 256$) stated they had not been shamed online before, and just over 17% indicated they were not sure ($n = 71$). For those participants who had been shamed online before and elaborated on their experiences, 25% ($n = 22$) reported being shamed about their appearance (e.g., weight, clothing) and 19% ($n = 17$) reflected they were shamed for having a differing opinion (e.g., views on vaccinations, veganism). Additionally, 7% ($n = 6$) disclosed they were shamed for their political views and 2% ($n = 2$) were shamed for their skin colour. Almost 15% ($n = 60$) of participants indicated they had shamed someone else online before, around 75% ($n = 308$) stated they had not shamed someone else online before, and just over 10% indicated they were uncertain ($n = 43$). Approximately 22% ($n = 13$) of these participants reported shaming another for holding opposing or differing views and 19% ($n = 12$) reflected shaming another over a discriminatory post made by that person (e.g., racist posts). Moreover, 10% ($n = 6$) of these participants reported shaming an individual due to their political views and 5% ($n = 3$) stated they shamed another over their appearance.

https://doi.org/10.1371/journal.pone.0279750

Recruitment methods included advertising on a social networking website (where snowballing occurred with other users sharing the post onto their pages; $n = 152$) and a university participant pool ($n = 259$). The opportunity to enter a prize draw was also offered to participants not within the university participant pool, with the only inclusion criteria being that participants were at least 14 years old (to meet the minimum age requirements of most social media platforms). An a-priori power analysis using G*Power 3.1 indicates that for multiple regression, a minimum sample of 114 participants was necessary for a medium effect size, with a power of .80 and significance level of .05 [24]. A medium effect size was chosen based on recent comparable research [e.g., 25]. Based on Kline's [26] criteria of 20 participants per free parameter, a minimum of 180 participants was necessary to adequately test the factor structure of the Online Shaming Scale. Hence, the obtained sample of 411 participants was deemed adequate to achieve meaningful analyses.

Measures

Criterion variabl. Previous measures of online shaming have generally used single or few items to assess online shaming behaviours, which impedes statistical testing and reduces variance [e.g., 27]. The Online Shaming Scale was developed by the authors for the purposes of this study and subsequent research to assess participant likelihood of engaging in online shaming. Based on principal axis factoring with promax rotation of an initial pool of 12 items with a randomised half of the sample ($n = 206$), followed by confirmatory factor analysis in EQS with the remaining half ($n = 205$), the final scale features nine items capturing various hypothetical shaming responses and opinions after a perceived social norm is violated online. All items assess level of agreement via a 7-point Likert-type scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). Scores are summed to produce a total online shaming score ranging between 9 and 63, with higher scores indicating a higher likelihood of engaging in online shaming. Total scores were also computed for the two subscales: 'online shaming intention' (items 3, 4, 7, 8, 10, 12) and 'online shaming perceived deservedness' (items 1, 6, 9), which were used separately in subsequent analyses. See S1 Appendix for further details pertaining to the development of this measure.

Predictor and scale control variables. Pre-existing measures for predictor variables (moral grandstanding, moral disengagement, emotional reactivity, empathy, social vigilantism, online disinhibition, machiavellianism, narcissism, and psychopathy) and the control variable social desirability are detailed in Table 3.6. All items in the Moral Grandstanding Scale [28] were adapted so that they only asked about moral beliefs instead of moral and political beliefs combined (e.g., in the item "I hope that my moral/political beliefs cause other people to want to share those beliefs", "/political" was removed). For the Perth Emotional Reactivity Scale- Short Form [29], only the items relating to negative emotional reactivity were used as positive emotional reactivity was not expected to be related to online shaming. Question 14 in the Social Vigilantism Scale [30, p.22] was adapted from "I frequently consider writing a letter to the editor" to "I frequently consider writing a product or service review" to reflect contemporary media. Question 1 in the Online Disinhibition Scale [31, p.256] was also adapted from "It is easier to connect with others through ICTs than talking in person" to "It is easier to connect with others online than talking in person" to reflect contemporary language.

Demographic variables. Single-item questions concerning demographics (age, gender, country of residence, religion, ethnicity, education attainment level, occupation, daily hours spent online, daily hours spent on social media, previously been shamed online before, and previously shamed someone else online before) were also collected to describe the sample and to be assessed as potential control variables.

Qualitative question. Participants also had the opportunity to respond to the optional open-ended qualitative question, "is there anything you would like to say about online shaming?".

https://doi.org/10.1371/journal.pone.0279750

https://doi.org/10.1371/journal.pone.0279750

Table 3.6 Predictor and control variable measures.

Construct/s	Measure	Factors	Items	Responses	Reversed items	Scoring	α	Example items
Moral grandstanding	Moral Grandstanding Scale [28]	2	10	1-7 (strongly disagree to strongly agree)	-	Mean scores from 1-7	.81-.95	"My moral beliefs should be inspiring to others"
Moral disengagement	Propensity to Morally Disengage Scale [32]	8	16	1-7 (strongly disagree to strongly agree)	-	Mean scores from 1-7	.88	"It's okay to gloss over certain facts to make your point"
Emotional Reactivity	The Perth Emotional Reactivity Scale- Short Form [29]	3	9	1-5 (very unlike me to very like me)	-	Total scores from 9-45	.76-.91	"I tend to get upset very easily"
Empathy	Basic Empathy Scale [33]	2	20	1-5 (strongly disagree to strongly agree}	1,6,7, 8, 13, 18,19,20	Total scores from 20-100	.85.91	"My friends' emotions don't affect me much"
Social vigilantism	Social Vigilantism Scale [30]	1	14	1-9 (disagree very strongly to agree very strongly)	-	Mean scores from 1-9	.81-.88	"I feel as if it is my duty to enlighten other people"
Online disinhibition	Online Disinhibition Scale [31]	2	11	0-3 (disagree to agree)	-	Total scores from 0-33	.81-.85	"I feel like a different person online"
Machiavellianism, narcissism, and psychopathy	The Short Dark Triad [34]	3	27	1-5 (disagree strongly to agree strongly)	2. 6. 8 (N) 2, 7 (P)	Mean scores for each factor from 1-5	.71.77	"It's not wise to tell your secrets" (M) "People see me as a natural leader" (N) "People often say I'm out of control" (P)
Social desirability	Marlowe-Crowne Social Desirability Scale- Adapted Version [35, 36]	1	13	True or false	1,2,4, 5, 7, 9, 11, 12	Total scores from 0-13	.74	"I sometimes try to get even rather than forgive"

Note. Measures presented here are self-report scales with higher scores indicating higher levels of each construct. All scoring instructions have been followed according to each measures' respective author guidelines, α = Cronbach's Alpha; M = machiavellianism; N = narcissism; P = psychopathy.

https://doi.org/10.1371/journal.pone.0279750.t001

Procedure

Following ethics approval by Curtin University's Human Research Ethics Committee (approval number: HRE2019-0697), the survey was piloted (i.e., a few individuals were asked to complete the survey first to ensure comprehension) before data collection took place (between January 23 and September 4, 2020) via a 30-minute online survey. Participants were first directed to a participant information sheet and consent form hosted by Qualtrics.com. After providing informed consent (by ticking a box) and completing validity questions to ensure they understood the contents of the survey, participants were presented with the demographic items, quantitative measures (which were randomised), and open-ended question. Afterwards, participants were debriefed and thanked for their time, and information regarding free online resources in the event of distress were provided. Upon completion, participants were redirected to either a page providing them the opportunity to enter a competition to win a gift voucher by submitting their email address, or a page where they could submit their enrolment details if they were from the university participant pool.

Data were downloaded from Qualtrics.com into IBM SPSS Statistics (v25) to be analysed, with qualitative responses transferred into a Microsoft Excel (v16.4) spreadsheet for analysis. While 500 individuals started the survey, 89 were removed from the quantitative analysis as they either did not complete any items for one or more scales, were found to be duplicates (i.e., identical IP addresses with the same answers) or were patterned responses. As such, data from 411 participants were retained for quantitative analysis and 153 qualitative responses for qualitative analysis. After screening and reverse coding, a missing values analysis indicated that data were missing completely at random, with no variable having missing data exceeding 0.5% (Little's MCAR test: $\chi 2(1130) = 1065.98$, $p = .913$). Consequently, the expectation-maximisation method was deemed appropriate for replacing missing values [37]. Mean and total scores were computed and categorical variables were dummy coded prior to analysis.

Results

Quantitative results

Table 3.7 provides descriptive statistics and correlations between predictor, criterion, control, and relevant demographic variables.

Hypothesis testing: To test the overarching quantitative hypothesis, that moral grandstanding, moral disengagement, emotional reactivity, social vigilantism, online disinhibition, machiavellianism, narcissism, and psychopathy will be positive predictors, and empathy will be a negative predictor of likelihood to engage in online shaming, two hierarchical multiple regression analyses (HMRAs) were completed (separate regressions for the online shaming subscales were conducted as differences between subscales in the correlation matrix were noted). The correlation matrix also revealed associations between online shaming and gender, daily online hours, daily social media hours, having shamed someone online before, and social desirability, and therefore these variables were used as additional control variables. All demographic control variables were entered into step one of both regressions, with social desirability entered as another control variable in step two, and the study predictors entered altogether in step three (in no particular order as there was no theoretically driven reason to do so). Assumptions underlying multiple regression were evaluated, with all relevant variables approximating normality except for the online shaming intention subscale, which approximated an L-shaped distribution and had heteroscedastic residuals [37]. Data transformation following guidelines by Tabachnick and Fidell [37] were conducted to remedy this, however this did not alter the overall findings or improve the skewness of the variable, so the original data were retained for analysis. See S2 Appendix for all normality plots. Univariate outliers were reduced to the most extreme non-outlier values plus one unit to maintain rank order [37], and no highly influential multivariate outliers were present. All other assumptions were met, and statistical significance for this hypothesis testing was evaluated at $\alpha = .05$.

https://doi.org/10.1371/journal.pone.0279750

Table 3.7 Descriptive statistics and correlations between predictor, criterion, control, and demographic variables ($N = 411$).

Variable	Descriptives			Correlations																		
	M	SD	α	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. OSS total	23.12	9.17	.84	-																			
2. OSS intention	13.20	6.87	.86	.93**	-																		
3. OSS PD	9.92	3.81	.65	.74**	.43**	-																	
4. Moral grandstanding [28]	3.53	0.81	.79	.43**	.42**	.28**	-																
5. Moral disengagement [32]	2.47	0.86	.89	.47**	.44**	.34**	.35**	-															
6. Emotional reactivity [29]	27.66	8.39	.91	.14**	.09	.19**	.08	.12*	-														
7. Empathy [33]	77.60	10.10	.89	-.24**	-.26**	-.11*	-.07	-.37**	.15**	-													
8. Social vigilantism [30]	5.17	1.10	.85	.45**	.42**	.33**	.58**	.42**	.15**	-.05	-												
9. Online disinhibition [31]	12.34	5.34	.76	.31**	.32**	.16**	.28**	.40**	.26**	-.11*	.33**	-											
10. Machiavellianism [34]	2.77	0.64	.79	.38**	.33**	.31**	.32**	.60**	.18**	-.25**	.45**	.41**	-										
11. Narcissism [34]	2.60	0.58	.72	.25**	.24**	.17**	.29**	.26**	-.06	-.18**	.41**	.15**	.36**	-									
12. Psychopathy [34]	2.04	0.60	.76	.43**	.45**	.24**	.21**	.57**	.19**	-.33**	.30**	.27**	.50**	.37**	-								
13. Social desirability [35, 36]	5.67	2.57	.64	-.13*	-.09	-.15**	-.10*	-.26**	-.35**	-.07	-.23**	-.22**	-.29**	-.10	-.26**	-							
14. Age	24.76	9.59	-	.04	.08	-.07	-.13**	-.22**	-.19**	<-.01	-.07	-.16**	-.09	-.05	-.12*	.17**	-						
15. Gender	-	-	-	-.13**	-.16**	-.04	-.02	-.23**	.12*	.28**	-.09	-.15**	_.15**	-.15**	-.27**	.04	.02	-					
16. Years of education	14.73	3.30	-	.01	.01	-.01	-.05	.11*	-.05	.05	-.03	-.15**	<.01	-.02	-.09	.05	.34**	.03	-				
17. Daily online hours	6.11	2.86	-	.12*	.11*	.09	.03	.12*	.16**	-.07	.07	.11*	.12*	.07	.15**	-.12*	-.18**	-.06	-.02	-			
18. Daily SM hours	3.16	2.14	-	.12*	.12*	.07	.17**	.15**	.21**	.10*	.19**	.21**	.11*	.10*	.10*	-.13**	-.26**	.13**	-.09	.55**	-		
19. Been OS	-	-	-	.03	.06	-.02	.05	-.08	.05	.15**	.06	<.01	-.11*	.11*	.04	-.12*	.09	.08	-.03	.02	.04	-	
20. OS someone	-	-	-	.28**	.29*	.15**	.20**	.12*	.06	.03	.17**	.06	.11*	.11*	.21**	-.17**	.11*	-.07	-.02	.08	.01	.27**	-

Note. Gender was coded as 0 = not women, 1 = women. Been online shamed and online shamed someone were coded as 0 = no or unsure, 1 = yes. *M* = mean; *SD* = standard deviation; α = Cronbach's Alpha; OSS = Online Shaming Scale; PD = perceived deservedness; SM = social media; OS = online shamed.

*$p < .05$.

**$p < .01$

https://doi.org/10.1371/journal.pone.0279750.t002

The first HMRA was conducted to assess the proportion of variance in online shaming intentions that could be accounted for by moral grandstanding, moral disengagement, emotional reactivity, empathy, social vigilantism, online disinhibition, machiavellianism, narcissism, and psychopathy (after controlling for gender, daily online hours, daily social media hours, having shamed someone online before, and social desirability). On step 1 of this HMRA, gender, daily online hours, daily social media hours, and having shamed someone online before collectively accounted for a significant 12% of the variance in online shaming intentions, $R^2 = .12$, adjusted $R^2 = .11$, $F\ (4, 406) = 14.03$, $p < .001$. On step 2, social desirability was added to the regression model, accounting for a non-significant additional < .01% of the variance in online shaming intentions, $\Delta R^2 = .00$, $\Delta F\ (1, 405) = .10$, $p = .748$. These five variables in combination still explained 12% of the variance in online shaming intentions, $R^2 = .12$, adjusted $R^2 = .11$, $F\ (1, 405) = 11.22$, $p < .001$. On step 3, moral grandstanding, moral disengagement, emotional reactivity, empathy, social vigilantism, online disinhibition, machiavellianism, narcissism, and psychopathy were added to the regression model, accounting for a significant additional 27% of the variance in online shaming intentions, $\Delta R^2 = .27$, $\Delta F\ (9, 396) = 19.60$, $p < .001$. All variables in combination explained 39% of the variance in online shaming intentions, $R^2 = .39$, adjusted $R^2 = .37$, $F\ (9, 396) = 18.27$, $p < .001$. This is a large effect by Cohen's [38] conventions ($f^2 = .64$).

The second HMRA was conducted to assess the proportion of variance in perceived deservedness of online shaming that could be accounted for by moral grandstanding, moral disengagement, emotional reactivity, empathy, social vigilantism, online disinhibition, machiavellianism, narcissism, and psychopathy (after controlling for gender, daily online hours, daily social media hours, and having shamed someone online before, and social desirability). On step 1 of this HMRA, gender, daily online hours, daily social media hours, and having shamed someone online before collectively accounted for a significant 3% of the variance in perceived deservedness of online shaming, $R^2 = .03$, adjusted $R^2 = .02$, $F\ (4, 406) = 3.27$, $p = .012$. On step 2, social desirability was added to the regression model, accounting for a significant additional 1% of variance in perceived deservedness of online shaming, $\Delta R^2 = .01$, $\Delta F\ (1, 405) = 5.56$, $p = .019$. These five variables in combination explained 4% of the variance in perceived deservedness of online shaming, $R^2 = .04$, adjusted $R^2 = .03$, $F\ (1, 405) = 3.76$, $p = .002$. On step 3, moral grandstanding, moral disengagement, emotional reactivity, empathy, social vigilantism, online disinhibition, machiavellianism, narcissism, and psychopathy were added to the regression model, accounting for a significant additional 16% of the variance in perceived deservedness of online shaming, $\Delta R^2 = .16$, $\Delta F\ (9, 396) = 8.70$, $p < .001$. All variables in combination explained 20% of the variance in perceived deservedness of online shaming, $R^2 = .20$, adjusted $R^2 = .17$, $F\ (9, 396) = 7.17$, $p < .001$. This is a medium effect by Cohen's [38] conventions ($f^2 = .25$). Regression coefficients, squared semi-partial correlations, and confidence intervals for both models are displayed in Table 3.8. Additionally, see S3 Appendix for a comparison between the current linear regressions and post hoc Ridge regressions as an additional robustness check of the current findings.

Qualitative findings

A mixed (conventional and summative) content analysis [39] was used to analyse the open-ended survey question "is there anything you would like to say about online shaming?". The conventional approach was used to describe emergent categories, with the summative approach used to quantify the frequency of responses within each category [39]. All participant responses were first read in their entirety by one researcher, before open coding was conducted with relevant meaningful responses noted and afterwards assigned a phrase or key word summarising each response. After two rounds of open coding, categories were created and then split, combined, or expanded upon where necessary. The summative aspect involved tallying the frequency with which the final categories occurred in the dataset, with words or phrases with interchangeable meanings also included within this count. To assess inter-rater reliability, approximately 10% of statements were classified as either in support of each category or against by two researchers using IBM SPSS Statistics (v25). Both coders have been trained in and have experience conducting qualitative analysis. An excellent level of inter-rater agreement ($\kappa = .75$) was found [40].

https://doi.org/10.1371/journal.pone.0279750

Table 3.8 Summary of regression coefficients and squared semi-partial correlations (sr^2) for two hierarchical regression models predicting online shaming intentions and perceived deservedness ($N = 411$).

	Predictors	OSS intentions					OSS perceived deservedness				
		Coefficient		95% CI Around *B*		sr^2	Coefficient		95% CI Around *B*		Sr^2
		B	β	LB	UB		*B*	β	LB	UB	
Step I											
	Gender	-2.47	-.16	-3.96**	-.98**	.02	-.27	-.03	-1.15	.60	<.01
	Daily online hours	.03	.01	-.23	-.98	<.01	.07	.05	-.09	.22	<.01
	Daily SM hours	.38	.12	.03*	.74*	<.01	.09	.05	-.12	.30	<.01
	OS someone	5.41	.28	3.63**	7.19**	.08	1.54	.14	.50**	2.59**	.02
Step 2											
	Gender	-2.46	-.15	-3.95**	-.97**	.02	-.23	-.03	-1.10	.63	<.01
	Daily online hours	.03	.01	-.23	.30	<.01	.06	.05	.09	.22	<.01
	Daily SM hours	.38	.12	.02*	.73*	<.01	.07	.04	-.14	.28	<.01
	OS someone	5.36	.28	3.56**	7.17**	.07	1.33	.12	.28*	2.39*	.02
	Social desirability	-.04	-.02	-.29	.21	<.01	-.17	-.12	-.32*	-.03*	.01
Step 3											
	Gender	<.01	<.01	-1.35	1.36	<.01	.30	.03	-.57	1.12	<.01
	Daily online hours	.07	.03	-.16	.30	<.01	.08	.06	-.07	.22	<.01
	Daily SM hours	<.01	<.01	-.32	.32	<.01	-.10	-.05	-.30	.11	<.01
	OS someone	3.51	.18	1.93**	5.09**	.03	.72	.07	-.29	1.73	<.01
	Social desirability	.26	.10	.03*	.49*	<01	.02	.01	-.13	.17	<.01
	Moral grandstanding	1.50	.18	.68**	2.32**	.02	.52	.11	<−.01	1.04	<.01
	Moral disengagement	.92	.11	.01*	1.83*	<.01	.88	.20	.30**	1.47**	.02
	Emotional reactivity	<−.01	<−.01	-.08	.07	<.01	.06	.14	.02*	.11*	.01
	Empathy	-.08	-.12	-.14*	-.02*	<.01	-.01	-.03	-.05	.03	<.01
	Social vigilantism	1.13	.18	.46**	1.80**	.02	.52	.15	.09*	.95*	.01
	Online disinhibition	.16	.12	.04**	.28**	.01	-.05	-.06	-.12	.03	<.01
	Machiavellianism	-.55	-.05	-1.69	.60	<.01	.59	.10	-.14	1.32	<.01

https://doi.org/10.1371/journal.pone.0279750

	Predictors	OSS intentions					OSS perceived deservedness				
		Coefficient		95% CI Around *B*		sr^2	Coefficient		95% CI Around *B*		Sr^2
		B	β	LB	UB		*B*	β	LB	UB	
	Narcissism	-.48	-.04	-1.57	.61	<.01	.02	<.01	-.68	.72	<.01
	Psychopathy	2.77	.24	1.57**	3.97**	.03	-.06	-.01	-.83	.71	<.01

Note. CI = confidence interval; OSS = Online Shaming Scale; *B* = unstandardised coefficient; β = standardised coefficient; LB = lower bound; UB = upper bound.

*$p < .05$.

^$p < .01$

https://doi.org/10.1371/journal.pone.0279750.t003

Holding people accountable, depicting online shaming as having destructive effects, and perceived anonymity afforded by the online world were the most frequent categories to emerge. Multiple participants also explored online shaming as a form of entertainment, as a phenomenon with multiple perspectives, and as often perpetrated by individuals coming from a place of shame or hurt. Some participants also regarded online shaming as now a social norm, as well as exploring perceived differences between public and private figures being shamed online. The complete findings of this content analysis are displayed in Table 3.9.

Table 3.9 Content analysis of what participants have to say about online shaming (*n* = 153).

Category	*n*	Description	Example quotes
Accountability	10	Online shaming is deemed acceptable and encouraged in certain settings when undertaken to hold an individual or group of individuals accountable and/or to raise awareness. Some participants appeared to view engaging in online shaming as serving the greater good, as it allows individuals to be informed and hold the perpetuator responsible for their actions online.	"I only participate in shaming behaviour if I know for a fact that they're acting abusive and needs to be held accountable for it"
Destructive effects	7	Participants referenced the harmful impacts of online shaming in a general sense but tended to avoid specific details. This may indicate that while there is a level of public awareness, people may not be aware of the specific, tangible consequences of online shaming.	"Its very dangerous and can cause alot of harm"
Anonymity	7	Anonymity provided through being online and the use of private accounts may allow people to feel more comfortable and safe to engage in online shaming. This sense of anonymity may afford people perceived power, which may embolden them to engage in online shaming. Participants directly contrasted online shaming to face-to-face bullying and suggested this would not occur in the latter setting.	"people feel they are anonymous online and feel safe to write evil hurtful things"

(*Continued*)

Table 3.9 Content analysis of what participants have to say about online shaming (n = 153). (*Continued*)

Category	*n*	Description	Example quotes
Entertainment	**6**	This category describes online shaming, including the use of memes, as a form of entertainment. Participants also articulated that online shaming and sharing memes may serve as a way to engage with friends and other community members.	"I only online shame them so that other people (who follow my private account) are aware of the things they say and we can laugh about it."
Two sides to every story	**6**	While there are two sides to every story, participants articulated that individuals should have the ability to express their opinion without being shamed. Online shaming and freedom of speech appears almost paradoxical; people expressing their opinion without worrying about being shamed is freedom of speech, but opposingly freedom of speech is also having the right to shame someone if you do not agree with them.	"There is always 2 sides to every story and everyone has the right to their own opinion."
"Hurt people hurt people"	4	Some participants referenced that being shamed online may cause a victim of online shaming to turn into a perpetrator. There is a belief that the perpetrator of online shaming may engage in this behaviour due to insecurities or past experiences with being shamed.	"It is a vicious cycle of people being online shamed and then going on to online shame others"
Social norm	3	Multiple participants referenced that online shaming has become a social norm.	"Online shaming is considered a social norm"
Public versus private figures	3	Participants reflected that individuals are held to different standards, depending on whether they are "public figures" or 'everyday' people. Public figures appear to be more readily online shamed, with some suggesting they are more deserving of this, due to having chosen to become a public figure or having a public platform.	"I am quick to call public figures out, as they have chosen a position of influence. With random people, I'm more restrictive."

Note. Single participant responses were able to be coded into multiple categories.

https://doi.org/10.1371/journal.pone.0279750.t004

Discussion

The main purpose of this study was to assess the utility of several currently untested predictors of online shaming, and given this research area is still in its infancy, we also sought to qualitatively capture general public opinions about online shaming. Quantitative results indicated that in combination, moral grandstanding, moral disengagement, emotional reactivity, empathy, social vigilantism, online disinhibition, machiavellianism, narcissism, and psychopathy accounted for approximately two fifths of the variance in intentions to engage in online shaming, and one fifth of the variance in perceived deservedness of online shaming (after controlling for gender, daily online hours, daily social media hours, having shamed someone online before, and social desirability). Some qualitative responses converged with quantitative findings regarding moral disengagement, empathy,

social vigilantism, and online disinhibition, whilst other qualitative responses suggested alternative or divergent understandings of online shaming. These findings are contextualised and explored in detail below.

Contextualising the current findings

Scores for the intentions subscale of the Online Shaming Scale indicated that on average, participants *disagreed* with the notion of intending to shame someone online. Scores for the perceived deservedness subscale were slightly higher, with participants on average only *somewhat disagreeing* that people who were shamed online deserved it. With the two subscales combined, participant total scores for online shaming fell between *disagree* and *slightly disagree*, indicating an overall trend towards not engaging in online shaming. Interestingly, the two subscales of the Online Shaming Scale only had a moderate positive correlation with each other, which combined with the slightly higher perceived deservedness scores, indicates that perhaps while some individuals may believe people to be deserving of being shamed online, they may be less willing to conduct the actual shaming themselves. From a measurement standpoint, this also supports the inclusion of both online shaming subscales rather than just one or the other.

Demographic data indicated that approximately one fifth of participants reported having been subjected to online shaming before, and slightly less than one fifth of participants reported having shamed someone else online before. In comparable research (e.g., exploring online harassment, cyberbullying), whilst some results indicate lower rates, many report higher [see 41–43]. The comparatively lower prevalence of online shaming in the current study may not be due to an actual lack of online shaming occurring, but as echoed previously within the literature [e.g., 1], participants may instead just not label their actions as online shaming. Instead, people may consider their actions to be a social deterrent and a mechanism to bolster accountability, a notion which was also salient within the current qualitative findings. Moreover, previous engagement with online shaming victimisation and perpetration were both only measured using a single item rather than measuring various types online shaming separately, indicating a lack of specificity which may also explain possibly understated rates of online shaming. The notion of online shaming now being considered a common and expected behaviour in society, which is reflected in both previous literature [e.g., 11] and the current qualitative findings, also serves to contrast these lower reports of online shaming. Another noteworthy finding is that prior online shaming had no significant relationship with any other demographic variable, which is interesting considering factors such as the gendered nature of online shaming (e.g., slut shaming, body shaming) often reported in previous literature [e.g., 15, 44]. There was also no significant relationship between online shaming victimisation and perpetration, which also contrasts related claims (i.e., the cybervictimisation to cyberbullying cycle often reported in cyberbullying literature) in similar studies [43], as well as contrasting participant perceptions offered in the current qualitative findings. These inconsistencies between the demographics in the current study and that of previous literature may simply be an indication of sample specific findings, or perhaps due to measurement differences between online shaming studies, or changing trends in online shaming. Regardless, this warrants further investigation into the relationship between demographic differences and online shaming in future research.

Whilst there were no significant associations between intentions to online shame and age, years of education, having been subjected to online shaming before, and social desirability, men and those who spent more hours online and on social media each week were slightly more likely to intend to engage in online shaming. There was also a moderate positive relationship between having shamed someone online before previously and future intentions to online shame. Although this was used as a control variable for subsequent hypothesis testing, previous online shaming actually accounted for the same unique variance in online shaming intentions as the strongest predictor, which echoes the age-old sentiment of the best predictor of future behaviour being past behaviour. As for the second online shaming subscale, there were also no significant associations between perceived deservedness of online shaming and age, gender, years

of education, daily hours online and on social media, and having been subjected to online shaming before. However, there was a small positive relationship between perceived deservedness of online shaming and having shamed someone online before previously, as well as a small negative correlation between perceived deservedness and social desirability. As for the predictors in this study, there were significant relationships (ranging from small to large) between all predictors and both online shaming subscales, with the exception of online shaming intentions and emotional reactivity. The predictive utility of these variables are further deconstructed below.

Moral predictors: Moral grandstanding and moral disengagement

Moral grandstanding and moral disengagement both had moderate positive correlations with both online shaming factors, indicating that participants with a stronger desire to appear morally credible to others and a propensity to convince themselves that ethical standards do not apply to them due to apparent extenuating factors were more likely to engage in online shaming and believe those who are shamed are deserving of this treatment. These findings are consistent with current theoretical musings relating to online shaming [e.g., 12, 13], the cyberbullying literature [e.g., 14], as well as some sentiments expressed by participants in the current qualitative findings. In the regression models, both moral predictors accounted for unique variance in intentions to online shame, however only moral disengagement (and not moral grandstanding) accounted for unique variance in perceived deservedness of online shaming (with moral disengagement being the strongest predictor of perceived deservedness).

Emotion predictors: Emotional reactivity and empathy

Emotional reactivity had a small positive correlation with perceived deservedness of online shaming, however it did not correlate significantly with intentions to online shame, suggesting that while participants who are more emotionally reactive may be more likely to believe people to be deserving of online shaming, this does not appear to impact their decision to actually engage in online shaming or not. This contradicts the notion of online shaming being a 'knee-jerk' reaction, as previously depicted in news media [15], which represents a potential discrepancy between the explanations hyper-publicised news stories provide for online shaming and what is actually occurring. Empathy however, was negatively correlated with both perceived deservedness and online shaming intentions (with the associations being small and small-to-medium in size, respectively), indicating empathy may be a protective factor against both believing people to be deserving of online shaming and intending to online shame. This coincides with studies on related online behaviours [e.g., 17], and is reinforced by participant concerns over the destructive effects of online shaming illustrated within the qualitative findings. In the regressions, emotional reactivity accounted for unique variance in perceived deservedness of online shaming, but not intentions to online shame. For empathy the opposite occurred, accounting for unique variance in intentions to online shame but not perceived deservedness.

Behavioural predictors: Social vigilantism and online disinhibition

Social vigilantism had medium positive correlations with both deservedness of online shaming and intentions to online shame, indicating that participants with a higher superiority of beliefs and perceived responsibility to propagate these beliefs were more likely to intend to engage in online shaming and believe victims of online shaming to be deserving. This aligns with literature that links online shaming to social vigilantism [e.g., 1, 18], with some current qualitative responses relating to accountability also converging with this notion. Online disinhibition had a moderate positive association with intentions to online shame, and a small positive correlation with perceived deservedness, demonstrating that those who feel more willing to express themselves freely on the internet and experience a lack of restraint online compared to in-person interactions are more likely to intend to online shame and perceive those who are shamed online to be deserving. Online disinhibition is often discussed in media reports as a potential contributor to online shaming [see 15], as well as within related literature [e.g., 19]. Online disinhibition also overlaps with the notions of anonymity presented

https://doi.org/10.1371/journal.pone.0279750

within the qualitative findings. In the regression models, social vigilantism was one of the largest unique predictors of online shaming intentions, and also accounted for unique variance in perceived deservedness. As for online disinhibition, it also accounted for unique variance in intentions to online shame, but not perceived deservedness.

Personality predictors: Machiavellianism, narcissism, and psychopathy

Machiavellianism had moderate positive correlations with both online shaming subscales, narcissism had small positive associations with both subscales, and psychopathy had a moderate to large positive relationship with online shaming intentions and a small positive relationship with perceived deservedness. This suggests participants who are more inclined to be deceitful, and exploitative to meet their goals (machiavellianism), who possess a sense of entitlement and a tendency for grandiose self-promoting (narcissism), and are characterised by a higher degree of impulsiveness, callousness, and selfishness (psychopathy), may be more likely to engage in online shaming and believe people to be deserving of online shaming. This also aligns with previous related research [e.g., 21] and has similarities to some notions described within reports (e.g., links to schadenfreude) in the media [15]. In the regression models, psychopathy was the largest unique predictor of intentions to online shame. However, no other personality predictor accounted for unique variance in either subscale, indicating psychopathic tendencies appear to be the most important dark personality trait assessed here for predicting online shaming.

Strengths, limitations, and future research

As the current paper is the first known study to empirically examine several currently untested predictors of online shaming, this research is strengthened by its originality. Considering empirical research in this domain is currently limited, another key strength was the inclusion of an array of potential predictors from various domains, and this study is also bolstered by the completion of pilot testing prior to data collection taking place. Given research [45] suggests effect sizes within social psychology research are typically small to medium (by Cohen's [38] conventions,) the medium and large effect sizes reported in this study make a worthwhile contribution to understanding why online shaming occurs. Other notable strengths include statistically controlling for social desirability, as well as advertising the study as 'online engagement' rather than 'online shaming' to minimise potential biases that may accompany the idea of shaming others. Sampling was not bound by any geographical restrictions (mirroring the globally accessible nature of the internet), which enhances the ecological validity of the current study. However, it should be noted that the current sample was skewed towards younger, Caucasian Australian university students who were employed and women. Future studies should endeavour to recruit more representative samples to gather a more generalisable understanding of online shaming engagement.

Given that, anecdotally, there are many differing understandings of what online shaming is, this study is also strengthened by the decision to provide participants with a definition of online shaming before completing questions relating to online shaming. However, given a specific shaming scenario was not used it is likely participants were envisioning different examples of online shaming when responding, in line with the availability heuristic [46]. Future exploration comparing different types of specific online shaming scenarios would be beneficial (e.g., featuring vignettes of different online shaming victims or varying social norm violations; or perhaps comparing online shaming across different social media platforms). The notion within the qualitative findings that people are held to differing standards when it comes to online shaming (e.g., being willing to shame public figures online but not everyday people) also substantiates this need to explore what conditions need to be met for people to be willing to engage in online shaming. It should also be noted that the moral choices individuals make in real life tend to be more self-serving than what hypothetical choices in research might capture [see 47]. Likewise, there is ongoing debate surrounding whether it is best to capture participant behaviours through self-reported intentions or via experimental means, as the use of hypothetical

scenarios has historically been criticised as an inaccurate way to measure behaviour [48]. Given this, future research using experimental methodologies (i.e., simulated online shaming scenarios where participants can respond as if the posts were real) may produce more accurate reflections of online shaming participation.

Although the qualitative strand of this study provides some interesting and novel findings that could not be captured by the quantitative strand alone, the use of a short open-ended question does not allow for any clarification of participant meanings, and the depth of responses is limited. For instance, while participants mentioned that online shaming had destructive effects, there was no opportunity to uncover what exactly these consequences were perceived to be. Additionally, it should be noted that not all participants in this study completed the open-ended question, meaning some self-selection bias may be present in the qualitative dataset (e.g., some participants who chose to take the additional time to provide a qualitative response may have done so due to feeling more strongly about certain views, having personal experience with online shaming, etc.). Semi-structured interviews exploring a range of perceptions and experiences of online shaming would allow for a deeper account of the complexities of this phenomenon.

Implications and conclusions

A key implication arising from this study is that the findings substantiate many previously posed but untested theoretical and media-driven explanations for online shaming, demonstrating the importance of multiple different factors in understanding this phenomenon. This provides academic and public discourse surrounding online shaming with a much-needed empirical basis and shift away from the current overreliance on purely theoretical and anecdotal speculation. This paper also offers a new validated measure to capture online shaming engagement, with two subscales which demonstrate that certain underlying mechanisms appear to be differentially responsible for a) individuals believing people to be deserving of being shamed online, and b) being willing to actually conduct online shaming.

As for practical implications, having an evidence-based understanding of what drives online shaming is an essential first step for policy makers to appropriately inform legislation. Additionally, when establishing formal guidelines and educational campaigns, the predictors tested in this study should be taken into consideration. For example, interventions could be designed (e.g., by social media companies themselves) to encourage empathy for others online and remind users to think before posting and that the individual being shamed is in fact a real person. While psychologists and other health professionals also need to have a comprehensive understanding when working with victims of online shaming, these findings may also be utilised to develop strategies to assist people and companies wanting to avoid being subjected to online shaming in the future, as well as provide the broader public a better understanding of why so many of us participate in online shaming ourselves.

Combined, this mixed-methods paper offers an original and significant contribution to the currently limited online shaming literature. This study has provided empirical support for several previously untested psychological explanations of shaming engagement in digital spaces, as well as qualitative insights into public opinions surrounding potential origins, concerns, and various other perceptual nuances of this phenomenon. Whilst the current findings can be utilised to inform public understandings, policy makers, educators, and health professionals, considerable further investigation into this research domain is still needed.

Supporting information

S1 Appendix. **Additional information for the development of the Online Shaming Scale.**
(DOCX)

S2 Appendix. **Normality plots for all scale measures.**
(DOCX)

https://doi.org/10.1371/journal.pone.0279750

S3 Appendix. **Linear and Ridge regressions using R.**
(DOCX)

Acknowledgments

We wish to extend our gratitude to Welber Marinovic for his invaluable assistance when conducting Ridge regressions as a robustness check for the current quantitative findings.

Author Contributions

Conceptualization: Shannon Raine Muir, Lynne Diane Roberts, Lorraine Sheridan.
Data curation: Shannon Raine Muir, Lynne Diane Roberts, Lorraine Sheridan.
Formal analysis: Shannon Raine Muir, Amy Ruth Coleman.
Funding acquisition: Shannon Raine Muir.
Investigation: Shannon Raine Muir, Lynne Diane Roberts, Lorraine Sheridan.
Methodology: Shannon Raine Muir, Lynne Diane Roberts, Lorraine Sheridan.
Project administration: Shannon Raine Muir.
Resources: Shannon Raine Muir.
Software: Shannon Raine Muir.
Supervision: Lynne Diane Roberts, Lorraine Sheridan.
Validation: Shannon Raine Muir, Lynne Diane Roberts, Lorraine Sheridan, Amy Ruth Coleman.
Visualization: Shannon Raine Muir, Lynne Diane Roberts, Lorraine Sheridan, Amy Ruth Coleman.
Writing – original draft: Shannon Raine Muir, Amy Ruth Coleman.
Writing – review & editing: Shannon Raine Muir, Lynne Diane Roberts, Lorraine Sheridan, Amy Ruth Coleman.

References

1. Skoric MM, Chua JPE, Liew MA, Wong KH, Yeo PJ. Online shaming in the Asian context: Community empowerment or civic vigilantism? *Surveill Soc*. 2010;8(2):181–199. https://doi.org/10.24908/ss.v8i2.3485
2. Cheung ASY. Revisiting privacy and dignity: Online shaming in the global e-village. *Laws*. 2014;3(1):301–326. https://doi.org/10.3390/laws3020301
3. Harris N, Maruna S. Shame, shaming and restorative justice: A critical appraisal. In: Sullivan D, Tifft L, eds. *Handbook of restorative justice*: *A global perspective*. Routledge; 2005:452–462.
4. Solove DJ. *The future of reputation*: *Gossip, rumor, and privacy on the internet*. Yale University Press; 2007.
5. Laidlaw EB. (2017). Online shaming and the right to privacy. *Laws*. 2017;6(1):1–26. https://doi.org/10.20944/preprints201612.0112.v1
6. Ingraham C, Reeves J. New media, new panics. *Crit Stud Media Commun*. 2016;33(5):455–467. https://doi.org/10.1080/15295036.2016.1227863
7. Massaro TM. The meanings of shame: Implications for legal reform. *Psychol Public Policy Law*. 1997;3(4):645–704. https://doi.org/10.1037/1076-8971.3.4.645
8. Ronson J. *So you've been publicly shamed*. Penguin; 2015.

9. Garvey SP. Can shaming punishments educate? *Univ Chic Law Rev*. 1998;65(3):733–794. https://doi.org/10.2307/1600299

10. Mann S. "Sousveillance": Inverse surveillance in multimedia imaging. Paper presented at: 12th Annual ACM International Conference on Multimedia; October 10–16, 2004; New York, United States. Accessed March 13, 2022. https://www.ics.uci.edu/~djp3/classes/2012_09_INF241/papers/sousveillance.pdf

11. Klonick K. Re-shaming the debate: Social norms, shame, and regulation in an Internet age. *Md L Rev*. 2016;75(4):1029–1065. https://digitalcommons.law.umaryland.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=3720&context=mlr

12. Gallardo KL. Taming the internet pitchfork mob: Online public shaming, the viral age, and the communications decency act. *Vand J Ent & Tech L*. 2017;19(3):721–746. https://scholarship.law.vanderbilt.edu/jetlaw/vol19/iss3/6

13. Spring VL, Cameron CD, Cikara M. The upside of outrage. *Trends Cogn Sci*. 2018;22(12):1067–1069. https://doi.org/10.1016/j.tics.2018.09.006 PMID:30340984

14. Kowalski RM, Giumetti G, Schroeder A, Lattanner M. Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychol Bull*. 2014;140(4):1073–1137. https://doi.org/10.1037/a0035618 PMID:24512111

15. Muir SR, Roberts LD, Sheridan LP. The portrayal of online shaming in contemporary online news media: A media framing analysis. *Comput Hum Behav Rep*. 2021;3(1):e100051. https://doi.org/10.1016/j.chbr.2020.100051

16. Mason L, Scrimin S, Tornatora MC, Zaccoletti S. Emotional reactivity and comprehension of multiple online texts. *Learn Individ Differ*. 2017;58(1):10–21. https://doi.org/10.1016/j.lindif.2017.07.002

17. Ang R, Goh D. Cyberbullying among adolescents: The role of affective and cognitive empathy and gender. *Child Psychiatry Hum Dev*. 2010;41(4):387–397. https://doi.org/10.1007/s10578-010-0176-3 PMID:20238160

18. Smallridge J, Wagner P, Crow JN. Understanding cyber-vigilantism: A conceptual framework. *J Theor & Philos Criminol*. 2016;8(1):57 70. https://search.proquest.com/docview/1787752058?accountid=10382

19. Chang LYC, Poon R. Internet vigilantism: Attitudes and experiences of university students toward cyber crowdsourcing in Hong Kong. *Int J Offender Ther Comp Criminol*. 2017;61(16):1912–1932. https://doi.org/10.1177/0306624X16639037 PMID:26992831

20. Dubrovsky VJ, Kiesler S, Sethna BN. The equalization phenomenon: Status effects in computer-mediated and face-to-face decision-making groups. *Hum Comput Interact*. 1991;6(2):119–146. https://doi.org/10.1207/s15327051hci0602_2

21. Kurek A, Jose PE, Stuart J. 'I did it for the LULZ': How the dark personality predicts online disinhibition and aggressive online behavior in adolescence. *Comput Hum Behav*. 2019;98(1):31–40. https://doi.org/10.1016/j.chb.2019.03.027

22. de Hert P, Papakonstantinou V. The new general data protection regulation: Still a sound system for the protection of individuals? *Comput Law Secur Rev*. 2016;32(2):179–194. https://doi.org/10.1016/j.clsr.2016.02.006

23. Frantziou E. Further developments in the Right to be Forgotten: The European Court of Justice's judgment in Case C-131/12, Google Spain, SL, Google Inc v Agencia Espanola de Proteccion de Datos. *Hum Rights Law Rev*. 2014;14(4):761–777. https://doi.org/10.1093/hrlr/ngu033

24. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods*. 2007;39(2):175–191. https://doi.org/10.3758/BF03193146 PMID:17695343

25. Ge X. Social media reduce users' moral sensitivity: Online shaming as a possible consequence. *Aggress Behav*. 2020;46(5):359–369. https://doi.org/10.1002/ab.21904 PMID:32488967

26. Kline RB. *Principles and practice of structural equation modelling*. 3rd ed. Guilford Press; 2005.

27. Blackwell L, Chen T, Schoenebeck S, Lampe C. When online harassment is perceived as justified. Paper presented at: Twelfth International AAAI Conference on Web and Social Media; June 25–28, 2015; Palo Alto, United States. Accessed March 13, 2022. http://www.lindsayblackwell.net/wp-content/uploads/2018/04/When-Online-Harassment-is-Perceived-as-Justified-ICWSM18.pdf

28. Grubbs JB, Warmke B, Tosi J, James SA, Campbell KW. Moral grandstanding in public discourse: Status-seeking motives as a potential explanatory mechanism in predicting conflict. *PLoS One*. 2019;14(10):e0223749. https://doi.org/10.1371/journal.pone.0223749 PMID:31618235

29. Preece D, Becerra R, Campitelli G. Assessing emotional reactivity: Psychometric properties of the Perth Emotional Reactivity Scale and the development of a short form. *J Pers Assess*. 2018;100(6):1–9. https://doi.org/10.1080/00223891.2018.1465430 PMID:29764211

30. Saucier DA, Webster RJ. Social vigilantism: Measuring individual differences in belief superiority and resistance to persuasion. *Pers Soc Psychol Bull*. 2010;36(1):19–32. https://doi.org/10.1177/0146167209346170 PMID:19776422

31. Udris R. Cyberbullying among high school students in Japan: Development and validation of the Online Disinhibition Scale. *Comput Hum Behav*. 2014;41(1):253–261. https://doi.org/10.1016/j.chb.2014.09.036

32. Moore C, Detert JR, Trevino LK, Baker VI, Mayer DM. Why employees do bad things: Moral disengagement and unethical organizational behavior. *Pers Psychol*. 2012;65(1):1–48. https://doi.org/10.1111/j.1744-6570.2011.01237.x

33. Jolliffe D, Farrington DP. Development and validation of the Basic Empathy Scale. *J Adolesc*. 2006;29(4):589–611. https://doi.org/10.1016/j.adolescence.2005.08.010 PMID:16198409

34. Jones DN, Paulhus DL. Introducing the Short Dark Triad (SD3): A brief measure of dark personality traits. *Assessment*. 2014;21(1):27–40. https://doi.org/10.1177/1073191113514105 PMID:24322012

35. Denton WH, Burleson BR. The Initiator Style Questionnaire: A scale to assess initiator tendency in couples. *Pers Relationsh*. 2007;14(2):245–268. https://doi.org/10.1111/j.1475-6811.2007.00153.x

36. Strahan R, Gerbasi K. Short, homogeneous versions of the Marlow-Crowne Social Desirability Scale. *J Clin Psychol*. 1972;28(2):191–193. https://doi.org/10.1002/1097-4679(197204)28:2<191::AIDJCLP2270280220>3.0.CO2-G

37. Tabachnick BG, Fidell LS. *Using multivariate statistics*. 5th ed. Pearson Education; 2007.

38. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. Erlbaum; 1988.

39. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res*. 2005;15(9):1277–1288. https://doi.org/10.1177/1049732305276687 PMID:16204405

40. Robson C. *Real work research*: *A resource for social scientists and practitioner-researchers*. 2nd ed. Blackwell; 2002.

41. Finn J. A survey of online harassment at a university campus. *J Interpers Violence*. 2004;19(4):468–483. https://doi.org/10.1177/0886260503262083 PMID:15038885

42. MacDonald CD, Roberts-Pittman B. Cyberbullying among college students: Prevalence and demographic differences. *Procedia Soc Behav Sci*. 2010;9(1):2003–2009. https://doi.org/10.1016/j.sbspro.2010.12.436

43. Xiao BS, Wong YM. Cyber-bullying among university students: An empirical investigation from the social cognitive perspective. *Int J Bus Inf*. 2013;8(1):34. https://doi.org/10.6702/IJBI.2013.8.1.2

44. Poole E. Hey girls, did you know: Slut-shaming on the Internet needs to stop. *U S F L Rev*. 2013; 48(1):221–260. https://ssrn.com/abstract=2483433

45. Lovakov A, Agadullina ER. Empirically derived guidelines for effect size interpretation in social psychology. *Eur J Soc Psychol*. 2021;51(3):485–504. https://doi.org/10.1002/ejsp.2752

46. Tversky A, Kahneman D. Judgment under uncertainty: Heuristics and biases. *Science*. 1974;185(4157):1124–1131. https://doi.org/10.1126/science.185.4157.1124 PMID:17835457

47. FeldmanHall O, Mobbs D, Evans D, Hiscox L, Navrady L, Dalgleish T. What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition*. 2012;123(3):434–441. https://doi.org/10.1016/j.cognition.2012.02.001 PMID:22405924

48. Yun GW, Park SY. Selective posting: Willingness to post a message online. *J Comput Mediat Comm*. 2011;16(2):201–227. https://doi.org/10.1111/j.1083-6101.2010.01533.x

3.10 Discussion Topics

The discussion theme for this chapter focuses on anonymity and free speech issues. Anonymity and freedom of speech are two of the most important democratic principles which help protect individual rights and diversity of voices while promoting open discussion. The two principles provide a basis for free expression but the relationship between them is complex and generates discussions about ethical boundaries and responsibilities. These topics can be explored in detail by considering the following discussion points:

1. Has the concept of freedom of speech changed as the Internet has grown?
2. Should online speech receive the same protections as physical speech or should new rules be established for digital platforms?
3. Social media platforms function as private domains or modern-day public forums. Do they have a responsibility to advance or suppress the right to free speech?
4. Social media algorithms determine which perspectives get visibility in online discussions to what extent?
5. Should social media platforms implement content moderation as a necessary step for maintaining platform order or does it infringe upon users' freedom of speech?
6. Fair content moderation requires which criteria and who should determine these guidelines?
7. The government needs to determine its proper role for online speech regulation while establishing methods to protect both free speech and public safety.
8. The government should decide whether to restrict online speech for public benefits or maintain complete non-intervention.
9. Anonymity serves as a key factor which supports free speech but it also leads to destructive conduct while diminishing personal responsibility.
10. Regulations that require online users to disclose their identities face opposition as they violate personal privacy rights.
11. Should ISPs maintain net neutrality through government regulation or should they retain full control over their network operations?
12. The freedom of speech together with the free internet information distribution depends heavily on net neutrality standards.

Utilize the knowledge you learn in this book together with your own experience, opinion and judgment will make your post much more attractive for discussion.

3.11 Projects

3.11.1 Navigating Violence in Media and School Safety: A Constitutional Approach

Abstract: The research project investigates the relationship between online violent content and video game violence and their impact on school shootings and evaluates safety measures that defend First Amendment rights. The project employs an interdisciplinary method which integrates psychological and sociological and legal perspectives to study violent digital content prevalence and its effects on youth behavior. The project will conclude with a set of proposals which minimize harmful effects of such content on school-aged individuals without violating constitutional freedoms.

Research Questions: Is there a significant correlation between exposure to violent media content and the propensity for school shootings? Is there any psychological theory that explains why young people become aggressive after watching violent media? What are the First Amendment limitations on regulating violent content and what are the consequences for public policy? Which evidence-based approaches could be established to reduce violent media content effects while protecting constitutional freedoms?

Methodology: Literature Review: A review of existing literature on the effects of media violence on youth and the legal framework of the First Amendment.

Expert Interviews: Brief interviews will be conducted with psychologists, legal experts and educators to gain an understanding of how media violence affects youth behavior and the legal system.

Survey: A survey will be created and distributed to a small group of high school students, teachers, and parents to collect preliminary data on the perception of media violence and its effects.

Analysis: Qualitative data from interviews and quantitative data from surveys will be analyzed to discover patterns and insights.

Expected Outcomes: A brief report will compile the information from the literature review, interviews and surveys. A set of preliminary recommendations for educators, parents, and policymakers that align with First Amendment protections. Further research recommendations that are based on the limitations and findings of this preliminary study.

Timeline:

- Week 1: Planning and Literature Review. This is the time when the project and its objectives are to be formulated. This is the time to start with a focused literature review on the relationship between violent media and school shootings, and the First Amendment limitations.
- Week 2: Interviews and Survey Development. Experts will be interviewed. A survey will be designed and prepared for distribution.
- Week 3: Data Collection. The surveys will be distributed and the responses will be collected. Start to analyze the data as it conies in.
- Week 4: Analysis and Reporting. This is the time to finish the data analysis. The results will be drawn and the final report will be written. Recommendations will be drawn from the research findings.

Conclusion: Given the limited time frame of this project, this will be a preliminary study that can guide a larger future project. The expected outcomes include a more accurate view of the current research, initial findings of what students and teachers think about violent media and some first amendment compliant recommendations for different stakeholders.

3.11.2 Anonymity versus Accountability: The Yelp Defamation Dilemma

Abstract: This four-week undergraduate research project examines the complex topic of hidden online assessments through a legal case involving a carpet cleaning service owner who sued seven Yelp reviewers who wrote defamatory content without revealing their identities. The research will analyze the reasons behind writing dishonest reviews while making two opposing cases about Yelp's responsibility to reveal user identities and the need to protect reviewer anonymity. The outcome will offer a fair evaluation of both the moral and legal aspects together with social effects of being anonymous online and being held accountable.

Research Questions: Online platforms feature what possible reasons do users post fake reviews? In which cases should online platforms reveal the identities of users who write defamatory content anonymously? The struggle between personal privacy rights and free speech rights exists against the right of business entities and individuals to defend their reputation from baseless accusations.

Methodology: Literature Review: A thorough investigation of legal decisions together with moral perspectives and social effects regarding digital anonymity and defamation.

Survey and Interviews: Survey participants will share their thoughts about online review behavior and interview professionals who specialize in digital law and ethics.

Mock Trial: A mock trial will enable the presentation of evidence to support or refute the need to reveal identities while using actual legal procedures along with ethical principles.

The analysis will examine data collection from both survey responses and mock trial arguments to determine the prevailing opinions and legal positions.

Expected Outcomes: This report will present the main factors that lead users to write misleading reviews on the internet. A debriefing session that includes solid evidence and counterarguments to support both sides of the issue regarding revealing identities of anonymous reviewers. A discussion about possible policies which online platforms should implement to protect user privacy and ensure accountability. The project provides a basis for future legal and ethical studies which could lead to policy or legislative development.

Timeline:

- Week 1: Project Planning and Literature Review
- The team will create a detailed project schedule before assigning participants into argument development teams. A comprehensive literature research on legal and ethical aspects of online anonymity and defamation begins at this stage.
- Week 2: Primary Research and Argument Development. Create a survey that will explore the public's views regarding anonymity and defamation within online review systems before distributing it to participants. Legal and ethical specialists must be interviewed for this project. Group one will develop arguments about reputation rights while group two will develop arguments about anonymity rights.
- Week 3: Mock Trial Preparation and Preliminary Analysis. Prepare for the mock trial by establishing roles, conducting legal research, and crafting opening and closing statements. The analysis of survey outcomes combined with the consolidation of interview data will help strengthen the arguments for both groups.
- Week 4: Mock Trial Execution and Reporting. Each group must present their case throughout the mock trial. The project report will incorporate both mock trial outcomes and primary research data to reach its final conclusions.

Conclusion: Through this project students will have the opportunity to examine real-world matters that connect legal frameworks with ethical considerations alongside digital communication platforms while developing their argumentation techniques and legal analytical abilities. The purpose of presenting both arguments equally will create an extensive knowledge base about the multiple dimensions of online anonymity versus accountability.

3.12 Exercises

1. How does the internet contribute to the spread of disinformation?
2. What is a common risk for individuals who express themselves on social media?
3. Why is it difficult for platforms to effectively review all content?
4. Why is cultural sensitivity important in content moderation?
5. What is a downside of relying solely on automated systems for content moderation?
6. What is the main challenge platforms face regarding freedom of speech?
7. Why is objectivity important in content moderation?
8. Why is transparency important in content moderation?
9. Why are content moderation decisions regarding newsworthy events controversial?
10. How does content moderation raise concerns about user privacy?
11. What role can the government play in promoting a safer online environment?
12. What concerns arise with government regulation of online speech?
13. Why is collaboration between governments and tech companies important in regulating online speech?
14. What challenges do governments face in regulating online speech globally?
15. What should be a priority when governments regulate online speech?
16. How does anonymity promote free expression?
17. What are negative consequences of anonymity in online speech?
18. What is a potential risk of requiring verified identities on social media?
19. What is pseudonymity, and how can it serve as a middle path?
20. What does net neutrality aim to ensure for internet users?
21. How does net neutrality protect freedom of speech?
22. What could happen if net neutrality is not regulated?
23. What is one argument for market-driven approaches to net neutrality?
24. What is a downside of government regulation in net neutrality?
25. What is throttling in the context of net neutrality?

Chapter

4

Privacy Risks and Principles

- The concept of privacy and how it has evolved in the digital age
- Examples of privacy violations and data breaches
- The impact of surveillance on individual privacy
- Strategies for protecting personal privacy online

4.1 The Evolution of Privacy in the Digital Age

Modern society has driven a significant shift in the way people think about privacy because of worldwide connections. The combination of technological progression and social media growth with data accumulation has disrupted basic privacy standards. This section investigates the transformation of privacy during the digital era while studying technological advancements' influence on individual privacy protection and information privacy maintenance [67, 71, 72, 114, 113, 137, 172, 176, 14, 63, 198, 102, 200].

Evolution of Privacy:

1. The Traditional Notion of Privacy: The idea of privacy serves as a basic human need throughout history. Throughout history people used physical boundaries to protect their personal areas as well as their intimate communications from unapproved access. The space where individuals could maintain privacy used to be limited to home environments and private conversations with trusted groups. The digital era has broadened privacy dimensions into complex systems.
2. The Digital Revolution and Its Implications: The internet together with digital devices brought fundamental changes to human communication patterns and information sharing practices. Email along with social media and e-commerce operations transformed personal data into a valuable asset. Through their online activities people share vast amounts of personal data either willingly or unintentionally while blurring the distinctions between what is private and public.
3. Social Media and the Erosion of Privacy: Social media platforms have reshaped how people understand privacy standards. These online platforms enable users to connect and express themselves but they create difficulties protecting their personal data. Users share their sensitive personal information and thoughts without recognizing the potential effects or how long this data will exist. The endless search for social validation through likes and followers has made people surrender their privacy through unintentional surveillance while becoming targets of advertising.
4. Data Collection and Surveillance: The modern digital environment has made personal data collection and analysis widespread throughout society. Multiple organizations including governments and corporations obtain massive data sets by monitoring browsing patterns and tracking user locations and performing online transactions. The gathered data provides beneficial results for personalized services and healthcare advancements along with research progress but creates problems related to data breaches and unauthorized access and potential misuse.

5. The Right to Privacy and Legal Implications: Modern legal frameworks face challenges because they have not been able to adapt to the evolving digital privacy environment. Various privacy laws across different jurisdictions fail to provide adequate protection because they differ from one another. Governments have created data protection laws such as GDPR in the European Union to define individual rights over personal information. The protection of privacy across all domains continues to be a difficult task for ongoing development.
6. Balancing Privacy and Convenience: People in the digital age face a recurring decision between protecting their privacy and getting convenience benefits. The provision of customized recommendations and targeted ads and smooth user experiences necessitates users to give away personal information. People today face a difficult task in achieving privacy-comvenience equilibrium because they need to measure the advantages of customized services against the dangers of data vulnerability and personal privacy diminishment.

In conclusion, the digital age has brought substantial changes to how privacy operates as a fundamental concept. The development of technology together with social media networks and data-focused operations transformed human understanding and experiences of privacy. Modern technology has provided extensive advantages yet introduced new privacy risks which threaten individual privacy rights. People must learn to navigate the changing world through awareness alongside consent and united efforts to build strong privacy-protecting laws which support digital innovation.

4.2 The Fourth Amendment and Changing Technology: Balancing Privacy and Security

As part of the Bill of Rights the Fourth Amendment protects citizens from unreasonable government searches and seizures by requiring judicial warrants with probable cause. The emergence of technological advancements brings new challenges which affect the interpretation of this amendment according to [114, 233, 67].

The Supreme Court has led the way in adjusting the Fourth Amendment to meet requirements of the digital age. The Carpenter ruling established privacy protections for new technologies which indicates the government needs proper process and oversight for searches and seizures of digital data. The courts have adopted this trend among others to strike a balance between privacy rights and law enforcement needs during the current technological evolution.

The Warren Court extended the exclusionary rule to cover every courtroom across the nation to prevent evidence obtained in violation of the Fourth Amendment from being used. In the Katz v. United States case the Supreme Court expanded Fourth Amendment protections to include electronic wiretaps while they adapted to new surveillance methods.

Technological advancements enable searches that were impossible before by allowing remote listening and infrared imaging and wireless tracking. The new surveillance technologies have raised questions about whether the Fourth Amendment requires reinterpretation to effectively address such advancements. The Riley v. California case illustrated how the Supreme Court faced fourth amendment protection challenges when evaluating mobile phone contents.

The Supreme Court faces challenges to the Fourth Amendment because it depends heavily on outdated precedents from before the modern technological era. Legal experts propose new methods that strictly follow the Fourth

Amendment's written language for protecting data and digital communications. The Libertas Institute demonstrates that privacy rights form the fundamental basis of the Fourth Amendment so the amendment must adapt to modern societal developments.

The Supreme Court applies the Fourth Amendment differently based on how technology affects government authority by strengthening protections during times of expanding government power but reducing them during times of threatened government authority.

In summary, the Fourth Amendment stands before a crucial decision regarding modern technological developments. This amendment protects physical spaces along with tangible items from government intrusions yet digital advancements require new definitions of "searches" and "seizures" together with privacy protection standards. The judicial system has started this process but the debate continues to affect both privacy rights and law enforcement capabilities.

4.3 Examples of Privacy Violations and Data Breaches: Lessons from the Digital Era

Data breaches and privacy violations have surged in the digital era to become major concerns about personal information protection. This presentation reviews specific privacy violations together with data breaches which demonstrate their effects on personal lives along with organizational activities and community dynamics. The frequency of such incidents demonstrates the need for strong personal data protection practices and demonstrates why privacy measures must be implemented in today's interconnected digital environment.

Examples:

1. Cambridge Analytica Scandal (2018): The Cambridge Analytica scandal exposed the full extent of personal data exploitation in political contexts. British consulting firm obtained access to the personal information of millions of Facebook users without their consent to develop targeted political advertising campaigns. Social media platform data exploitation led to major privacy and data ethics concerns along with democratic processes manipulation issues.
2. Equifax Data Breach (2017): In 2017 Equifax became one of the biggest credit reporting agencies in the United States when a massive data breach occurred. The hacking incident used an Equifax system weakness to access personal information belonging to about 147 million consumers. The data breach exposed highly sensitive information which included Social Security numbers and birth dates along with addresses thus putting affected people at risk of identity theft and financial fraud. This incident demonstrated the necessity of strong cybersecurity measures while raising doubts about how organizations should protect customer data.
3. The online business Yahoo suffered from two major data breaches throughout 2013 and 2014. User accounts containing names email addresses and hashed passwords were compromised in two data breaches that affected 3 billion Yahoo account holders. The incident made user data publicly accessible while simultaneously ruining Yahoo's image and diminishing customer trust. The incident showcases the enduring effects data breaches have on user trust as well as the critical role of proper security measures.
4. Marriott International Data Breach (2018): Marriott International announced in 2018 that their guest records were compromised in what became the largest data breach in history affecting 500 million people. The data leak exposed both personal names and addresses and passport details and payment card details. The incident showed how weak security measures affect both hospitality businesses and individual privacy along with financial safety. The incident reinforced the importance of organizations implementing effective cybersecurity practices to protect their customer information.
5. Google Street View Wi-Fi Collection (2010): Google Street View cars accidentally intercepted personal data from unprotected Wi-Fi networks while they gathered information for their mapping program. The data included emails together with passwords and browsing histories which raised

issues regarding unauthorized access and privacy violations. The data breach prompted worldwide regulatory investigations which ended with organizations facing penalties through legal settlements. The incident demonstrated why consent and transparency along with accountability need to be fundamental when processing personal information.

In conclusion, the digital era has witnessed an alarming increase in data breaches alongside privacy violations which threaten both personal and business entities. Multiple real-life incidents presented in this session reveal extensive consequences of such events which result in individual data exposure combined with trust breakdown and financial as well as reputational consequences. The ongoing technological development requires individuals together with organizations and policymakers to establish privacy protection measures and execute robust cybersecurity protocols and create specific regulations which defend personal data. The examples provide valuable lessons that lead us toward building a safer digital future which respects privacy.

4.4 The Impact of Surveillance on Individual Privacy

The development of sophisticated surveillance technology during the digital era has caused substantial effects on personal privacy. The extensive deployment of surveillance systems operated by both government and private companies creates privacy concerns about individual privacy rights while generating potential social effects for people and communities. The examination of surveillance effects on personal privacy investigates the consequences that affect both personal autonomy and societal interactions as well as the relationship between security measures and constitutional freedoms.

Impacts:

1. Loss of Anonymity: The implementation of surveillance technology including CCTV cameras and facial recognition systems and digital tracking tools has reduced public space anonymity for individuals. The reduction of anonymity creates various negative effects because people experience ongoing monitoring that leads to both internalized monitoring and restricted personal freedoms of expression.
2. Chilling Effect on Free Speech and Expression: People tend to restrict their free speech and expression because they understand they are under constant surveillance. People who know their activities are monitored tend to avoid sharing political views or participating in activism and exploring new ideas. Self-censorship harms democratic values by blocking the essential free exchange of ideas which drives societal advancement.
3. Impact on Personal Relationships and Intimacy: The installation of surveillance technology creates obstacles that prevent people from developing close relationships and maintaining intimate bonds with others. People tend to become distrustful when they understand their communications through both online and offline channels may be monitored. People avoid revealing their intimate thoughts which weakens human connection and destroys the privacy needed to build authentic emotional bonds.
4. Psychological Impact: People who experience constant surveillance tend to experience negative psychological effects. The permanent awareness of being watched generates increased stress levels along with anxiety and continuous alertness. Surveillance systems cause damage to mental health and general well-being while restricting personal freedom and the pursuit of authentic living.
5. Potential for Abuse and Discrimination: The use of surveillance technologies presents opportunities for their improper use that result in discriminatory practices and violations of personal rights. The implementation of biased algorithms and facial recognition systems results in unfair targeting of minority communities while sustaining existing systemic biases. Social profiling emerges when

authorities monitor people based on their racial or religious background or political affiliations which leads to violations of civil rights.

6. Balancing Security and Privacy: Security programs need to maintain an appropriate equilibrium between their protective functions and individual privacy rights. Effective legal systems combined with monitoring processes and protective mechanisms must exist to guarantee surveillance methods are essential and fair and follow ethical guidelines.

In conclusion, surveillance technology creates complex and multifaceted consequences for personal privacy. Surveillance technology advancement requires society to evaluate its effects on personal freedom alongside social dynamics and societal structures. Security requirements must harmonize with individual privacy protection measures to ensure both civil liberties and democratic principles and personal freedoms remain intact. The combination of open dialogue and transparent practices along with well-defined rules enables society to handle surveillance issues while defending the basic right to privacy in modern times.

4.5 Strategies for Protecting Personal Privacy Online

The digital age demands complete protection of online personal privacy because data collection methods continue to expand. People who protect their sensitive information and control their personal data reduce their exposure to privacy violations and identity theft risks. This session investigates multiple approaches to online privacy protection that allow people to independently defend their digital identity and maintain control of their personal data.

Strategies:

1. Every online account requires a strong password which should be unique from others. Choose passwords that combine upper and lowercase letters together with numbers and special characters instead of selecting easily discoverable or repeated options. A trustworthy password manager provides a secure platform to generate and store passwords.
2. Two-Factor Authentication (2FA): Enable two-factor authentication whenever possible. The additional security feature in 2FA requires users to verify their identity through a text message or email or authenticator application. The second verification process makes online account access highly challenging even for unauthorized users who possess the password.
3. Software updates and patches for operating systems and applications together with software should always be maintained to protect online privacy. Software updates contain essential security patches which fix system weaknesses. Regular software updates for devices along with their applications reduce the chance hackers or malware can exploit vulnerabilities.
4. Learn about the privacy options and authorization levels of all online services which you access. Examine privacy options to restrict your personal data exposure to both public viewing and third-party sharing. Review the permissions granted to applications to make sure they match your comfort standards for data access and usage.
5. People should use encryption technologies to safeguard their important information during online transfers. Look for websites and services that implement SSL or TLS security protocols because they display "https://" at the start of their URLs. People should select email services and messaging applications which use end-to-end encryption so content becomes accessible only to its intended recipients.
6. When posting online you should be aware of what data you release and control how much personal information you expose to the public. Social media users should avoid sharing personal information

and limit the amount of personal details they give to websites and forms. Users should evaluate their social media privacy settings to determine how much information they want shared publicly.

7. Virtual Private Networks (VPNs) protect internet traffic through encryption while hiding IP addresses to provide anonymity. VPNs provide essential security protection when using public Wi-Fi networks because they shield your online activities from potential eavesdroppers.
8. Never accept or click on links from unfamiliar emails or open attachments from unknown senders. Phishing attempts and scams lead to identity theft while also putting personal information at risk of compromise. Check the authenticity of all communications before submitting sensitive information to unknown parties.

In conclusion, the protection of individual privacy in the digital environment requires constant attention throughout the modern era. Strong passwords combined with two-factor authentication along with software updates and careful data sharing practices enable people to strengthen their digital privacy and minimize identity theft threats. Knowledge about ongoing privacy issues together with safe practices and active digital identity management remain vital for online safety. People who implement these measures will achieve digital confidence while preserving their personal data integrity.

4.6 Laws and Regulations for Protecting Privacy

The protection of privacy stands as a core human right which safeguards individual autonomy and dignity while securing sensitive information. The digital world's borderless data flow requires comprehensive laws which protect people from privacy violations during personal data collection and processing and sharing across international borders. This session examines privacy laws and regulations together with their development and their function in defending personal data throughout the digital period [71, 72, 113, 137, 172].

Privacy laws and regulations exist because societies acknowledge privacy as a basic human right. Various countries have established this right by writing it into their constitutional laws and legal systems. The United States Fourth Amendment safeguards against illegal searches and seizures but the European Union has established the General Data Protection Regulation (GDPR) as one of the most extensive privacy regulations worldwide.

Modern technology has revolutionized the way personal information gets collected and processed. The combination of online interactions and social media with e-commerce operations and data-driven technologies generates extensive amounts of personal data. Modern privacy legislation needs to incorporate strong protections which address digital age-specific privacy challenges.

Key Privacy Regulations:

1. The GDPR which was introduced in 2018 provides EU citizens with expanded control of their personal information. All organizations under GDPR must collect consent explicitly from individuals and build proper data protection systems and disclose data breaches as soon as possible.
2. The Federal Trade Commission (FTC) within the United States implements consumer privacy rules through its enforcement of unfair and deceptive data practices. The Children's Online Privacy Protection Act (COPPA) increases protections for kids under 13. At least 13 states adopted consumer data privacy legislation in 2021 while California serves as a leader with its thorough privacy regulations. The California Consumer Privacy Act (CCPA) became effective in 2020 and offers California residents particular rights regarding their personal information such as data collection information and the ability to request data removal and the right to choose data sharing.

3. Health Insurance Portability and Accountability Act (HIPAA) regulates health information privacy and security protection in the United States to guarantee medical data protection and controlled sharing.

Principles of Privacy Laws:

The main principles which govern privacy laws include the following elements:

1. Before data collection begins individuals must provide their consent for processing their information.
2. Data must be used exclusively for the functions established when data was initially collected.
3. Data should only include the minimum amount needed for the stated purpose of collection.
4. Security: Implement measures to protect data from breaches and unauthorized access.
5. Transparency: Inform individuals about data practices and allow them to access their own data.
6. The responsibility falls upon organizations to demonstrate both data protection compliance and to address any data breaches that occur.

The implementation of privacy laws encounters multiple barriers during their enforcement process. The worldwide nature of data transfer combined with conflicting legal domains creates obstacles in enforcement processes. Critics maintain that privacy regulations have not adapted to recent technological advancements and changing operational practices. The current legislative body faces ongoing debates regarding data encryption doors, social media data collection methods, AI algorithm oversight and online advertisement procedures. The FTC together with state attorneys general face difficulties in enforcing compliance while selecting violators for prosecution.

Modern privacy challenges arise from the development of emerging technologies including facial recognition systems along with biometric features and artificial intelligence systems. Privacy laws must evolve to include specific provisions for managing new technologies that spread across public spaces as these technologies become more widespread.

In conclusion, privacy laws together with regulatory frameworks stand as essential instruments which defend personal information while maintaining individual rights throughout our contemporary digital environment. These laws establish vital requirements for responsible data collection and use that honors both personal freedom and human dignity. Technology advancement demands privacy laws to transform so they handle new challenges effectively while balancing innovation with personal privacy protection. These legal frameworks guide the digital world's complex privacy terrain to build trust and accountability throughout the modern data age.

4.7 Case Study 1

RESEARCH ARTICLE

Intent to purchase IoT home security devices: Fear vs privacy

Joey F. George *, Rui Chen, Lingyao Yuan

ISBA Department, Ivy College of Business, Iowa State University, Ames, IA, United States of America

** jfgeorge@iastate.edu*

Abstract

The Internet of Things (IoT) is a widely hyped concept, with its focus on the connection of smart devices to the Internet rather than on people. IoT for consumers is often called the smart home market, and a large part of that market consists of home security devices. Consumers are often motivated to purchase smart home security devices to prevent burglaries, which they fear may lead to damage to their property or threats to their families. However, they also understand that IoT home security devices may be a threat to the privacy of their personal information. To determine the relative roles of fear and privacy concerns in the decision to purchase IoT home security devices, we conducted a survey of American consumers. We used the Theory of Reasoned Action as the theoretical basis for the study. We found that fear positively affected consumer attitudes toward purchasing smart home security devices, while concerns about privacy negatively affected attitudes. We found that attitudes toward purchase, the opinions of important others, and experience with burglaries all affected intent to purchase. We also found that the relationship between privacy concerns and intent to purchase is completely mediated by attitudes, while fear has both direct and indirect effects on intent.

Introduction

The Internet of Things (IoT), with its focus on connecting devices instead of people via the Internet, is one of the most hyped information systems topics of the last decade [1]. IoT for consumers is often referred to as the 'smart home' market and includes such things as smart televisions, smart speakers (e.g. Alexa), thermostats, lighting controls, video doorbells, and security cameras. The global market for the smart home industry is projected to be worth over $150 billion USD by 2023, while the US market, with 45 million smart devices in 2019, will be worth $42 billion [2,3]. While most consumers (84%) prize the smart home's convenience, 37% purchased IoT devices for home and family security. The global market for smart home security devices alone was $2.14 billion USD in 2018, and it is expected to increase to an estimated $5.05 billion by 2025 [4]. Certain smart security devices are popular–the most common home security device is a camera, and smart doorbells rank third [3]. Unit sales for cameras and smart doorbells increased by 125% between February 2017 and February 2018, and it is estimated that Ring, the Amazon owned maker of video doorbells and related security devices, sold more that 400,000 units in December 2019 alone [5,6].

Smart home security marketing mentions many reasons why a consumer would want to invest in such devices, but the marketing almost always appeals to fear. For example, safewise [7], an independent review site, answers the question "Do I Really Need a Security System?" with:

https://doi.org/10.1371/journal.pone.0257601 Used with CC 4.0 license

According to Alarm System Report, over 2 million home burglaries are reported in the United States each year and, on average, a burglary of a home in the U.S. occurs every 13 seconds-or about four burglaries a minute, 240 an hour, and nearly 6,000 a day.

Reolink, a vendor of home security cameras, lists in a blog 10 reasons why consumers need a home security system [8]. The first reason is "to secure your family and property." The post then cites the same statistics about burglaries that appear on the safewise site quoted above. They go on to say that "FBI burglary rates of homes state that 1 in 3 homes without a security system will fall victim to a burglary as compared to 1 in 250 homes that do have a security system," followed by videos–caught on Reolink cameras–of burglars attempting to break into a home and a car.

Yet consumers recognize there is a potential tradeoff between security and privacy. In a recent survey of over 1000 US consumers, 50.2% were most concerned about protecting their homes, but an almost equal proportion, 49.8%, were most concerned about protecting their personal data [3]. For example, Ring doorbells have become a privacy concern, with documented cases of smart doorbells leaking WiFi login credentials, with their known vulnerability to hackers, and the company's practice of sharing users' videos with law enforcement [9]. Perhaps the best-known case of a privacy breach from the use of a Ring product occurred in Mississippi in December 2019. An unknown man gained access to a Ring camera and watched an eight-year-old girl, asking her at one point to be his best friend, claiming he was Santa Claus [10].

To investigate the respective roles of fear of burglary and concerns over data privacy in the decision to purchase smart home security devices, we conducted a survey of potential consumers of such devices. We used the Theory of Reasoned Action [11] to form the theoretical framework for our study. We found that beliefs about fear affected the attitude toward purchasing the most, but beliefs about privacy also played a role. Attitude toward purchase and the views of important others affected purchasing intent about equally. Whether or not someone had personal experience with burglary also positively affected the intention to purchase smart home security devices.

In the next section, we review the literature and present the theoretical foundation for our study. We then describe our research design, data collection, and analysis. We end with a discussion of our findings and their implications.

Theory and Literature

The Theory of Reasoned Action (TRA) is a well-known and widely used theory for explaining individual intent to engage in a particular behavior [11,12]. Attitude toward the target behavior and subjective norms about engaging in the behavior are thought to influence intent. According to TRA, an individual's performance of a certain behavior is determined by his or her intent to perform that behavior. Intent is informed by attitudes and subjective norms. According to Ajzen [13], attitudes are informed by beliefs about the behavior, and norms are informed by normative beliefs and motivation to comply. Specifically, attitude toward a behavior is shaped by behavioral beliefs, readily accessible in the memory of a decision maker, which refer to the subjective evaluations of those consequences following performing the behavior. Subjective norms are under the influence of normative beliefs that concern the perceived expectations of referent individuals or groups, for which the decision maker is motivated to comply. As a general theory, TRA does not specify the particular beliefs that are associated with any particular behavior, nor does it specify referent individuals or groups. Both of these tasks are left to the researcher and are dependent on the nature of the behavior in question.

TRA has a long history of use in the social sciences. For example, a 1999 meta-analysis of 67 studies on the intent to use condoms found that attitudes toward condom use and subjective norms demonstrated medium to strong effect sizes [14]. TRA has been particularly useful as a framework for explaining the intent to

https://doi.org/10.1371/journal.pone.0257601

purchase consumer goods. Recent examples include the intent to buy a Volkswagen [15]; environmentally friendly (green) products [16]; "pioneer" products (those produced by a first mover firm in a market) [17]; "functional" yogurt (a functional food is one to which nutritional value has been added) [18]; "upcycled" products (upcycled products create value out of old or discarded materials) [19]; foreign goods [20]; and participation in online fashion renting [21]. Because it is functional, relatively simple, and reliable, we use TRA as the framework for our investigation into the relative influence of fear and privacy concerns on the intent to purchase IoT home security devices (Fig 4.1).

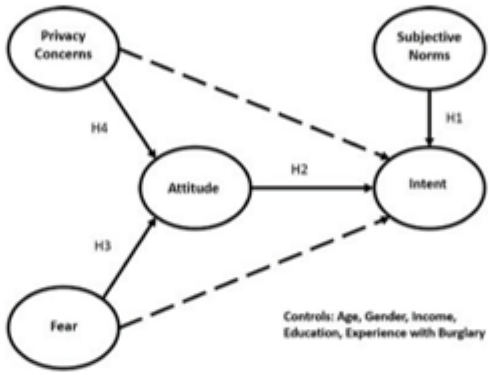


Fig 4.1 TRA model of intent to purchase IoT home security devices.
https://doi.org/10.1371/journal.pone.0257601.g001

Across a wide range of recent papers about the intent to purchase IoT devices, subjective norms and attitude were key factors in explaining intent [22–26]. Where subjective norms were investigated, the important others identified varied with context, as would be expected. Important others listed in these studies included retailers in general, competitors, and the media [22]; the generic "people important to me" [23]; and the generic "social circle" [25]. Where beliefs about purchasing were identified, both security [25,27–29] and privacy concerns [24–26,28] were key factors. In general, security concerns positively affected intent, while privacy concerns negatively affected it. Where included, price value either positively affected intent [25] or it had no effect [29].

Given these findings, and given the practical considerations for focusing on security and privacy concerns (mentioned earlier), we included both in the model as beliefs informing the attitude toward purchasing. Given all the different ways security concerns were operationalized, we decided to use fear as a surrogate. We did not include price value, given its general absence in other studies (in these IoT purchasing papers we reviewed, price value was not included in [22–24,26,27,30], the contradictory outcomes where it was included (no effect on intent in [29], but a positive effect on intent in [25]), and our inability to find or develop a suitable measure for the cost of smart home security devices. To cast a wide net for influential others, we decided to use friends, neighbors, family, and co-workers as important others. For control variables, we included age, gender, income, education, and experience with burglary.

Since its development in the 1970s, in many different studies across many different behaviors, the strong relationships between subjective norms and intent, and between attitude and intent, have become standard and expected. We therefore predict that:

> ***H1***: *Subjective norms about purchasing a home security IoT device positively influence the intent to purchase such a device.*
>
> ***H2***: *Positive attitudes towards the intent to purchase a home security IoT device positively influence the intent to purchase such a device*

Fear is an important construct in the literatures on how people respond to threats and on security, going back at least to its implicit inclusion in Protection Motivation Theory [31]. Rogers referred to fear as a "relational construct, aroused in response to a situation that is judged as dangerous and toward which protective action is taken" ([31], p. 96). In the case of deciding to purchase IoT home security devices, the situation that is judged as dangerous is having one's home burglarized, potentially damaging one's property and/or harming one's family. The protective action is purchasing and installing home security equipment. In the information systems security literature, fear has become an explicit construct [32]. The fear appeal literature suggests that fear is responsible for attitude change [31], so we would expect those with the most fear to be the most attracted to IoT home security devices.

> ***H3***: *Fears over home security will positively affect attitudes toward purchasing an IoT home security device.*

Our other key belief informing attitude toward purchasing, privacy concerns, has been important in studies of consumer purchasing intention and behavior for decades [33–35]. IoT industry and government agencies such as the US FTC have all stressed the privacy issue when guiding consumers about their purchasing decisions [36]. A systematic review of the behavioral literature on privacy and e-commerce [37] found eight different themes. Three of these themes are directly related to IoT devices and privacy: 1) e-commerce benefits and consumer privacy, 2) privacy-related risks and concerns, and 3) information control and privacy. Related to the first theme is the idea that consumers are likely to give up some degree of privacy if they believe they are receiving a fair exchange. In this case, the exchange would be between security and privacy. For the second concern, perceived risk is driven by the likelihood of negative consequences. Fortes and Rita [38] found that privacy concerns were related to increased perceptions of risk, which affected attitudes toward engaging in e-commerce. Regarding the third concern, control over private information has been prominent in the privacy literature. Many times control is considered to be a contextual matter rather than a personal disposition. Given a focus on fair exchanges, risk and control (or lack of it), consumers with privacy concerns related to information technology tend to be less likely to purchase technologies that have the potential to provide private information to unwanted others. Unlike traditionally "dumb" devices, IoT devices are equipped with the ability to collect personal data such as personal identities, geolocations, habits, and physical conditions. When consumers are aware of the potential risks to the data collected by their IoT devices, they may be less attracted to such devices.

> ***H4*****:** *Beliefs about the perceived privacy risks of IoT home security devices will negatively affect attitudes toward purchasing them.*

A central principle of TRA is that the influence of beliefs on intent are mediated by attitudes toward the behavior in question. Direct paths from the beliefs to intent are typically not tested. We have added those direct paths in Fig 4.1 (dotted lines). After we test the TRA structural model, we will test it a second time with the direct paths included, to determine the extent to which beliefs about purchasing smart home security devices are mediated by attitudes towards purchasing such devices in their effect on intent to purchase. We did not suggest a hypothesis regarding the direct paths, as we have no basis from the literature for doing so.

Research Method & Results

Measurement

To test our models (Fig 4.1), we searched for an existing survey instrument that measured all of the relevant constructs. As none existed, we created our own instrument, beginning with a search for pre-existing measurement scales for as many constructs as possible. TRA does not specify specific behaviors or referent groups, so standard TRA measures must be modified for the behavior and referent groups under study. Here we needed to find and modify suitable measures for key aspects of TRA, including intent, attitude, and

https://doi.org/10.1371/journal.pone.0257601

subjective norms. Measures for attitude and subjective norms were taken from [33], which was based largely on [39], and modified for the behavior in question, adopting Internet-of-Things home security devices. The scale for fear came from [40] and was modified for fear of having one's home burglarized. Privacy risk measures were derived from [41] and modified to capture beliefs about the risks to privacy from adopting IoT home security devices. Measures for intent were developed by the authors, as we were unable to find any suitable scales that measured the intent to adopt IoT devices for home security. All items, except those for 'attitude,' were measured using a 7-point Likert scale, from '1' for 'strongly disagree' to '7' for 'strongly agree.' The four 'attitude' items used semantic differential seven-point scales, ranging from bad to good, foolish to wise, dislike to like, and unpleasant to pleasant. We used the question of "Has your home or that of someone you know been burglarized?" to capture the burglary experience of consumers to be surveyed. This measure was developed by the authors. It was measured on a five-point scale: 1 = in the past year; 2 = in the past two years; 3 = in the past three to five years; 4 = more than five years ago; 5 = never burglarized. Finally, we used standard measures for the other control variables: age, gender, education and income.

During the development of the survey instrument, the authors went through several rounds of revision. The original versions of the instrument contained multiple measures of some constructs. All of the items to be used in the instrument were subjected to a sort procedure conducted individually by three doctoral students and one professor. Each item was printed on a separate piece of paper, and the judges were asked to sort them into however many piles they thought were appropriate. They were then asked to name each pile. There was considerable overlap among the judges, but there was no complete consensus. Some judges had more piles of items than others, and the distribution of items across piles differed by judge. The areas where they did not agree pointed to potential problems. The authors revised the instrument, based on the feedback from the sorting procedure, until the instrument was stable. The authors made several distinct changes in the instrument: 1) They changed the measure for intent, dropping two items and adding three new ones; 2) They added income and education to the demographic variables; and 3) They added the item about whether or not the respondent's home had been burglarized.

The revised instrument was then validated, beginning with a pilot study with 100 participants. The participants were recruited by Cint, a firm that provides access to over 40 million consumers worldwide, for a cost. To qualify for the survey, potential respondents had to indicate that they were not current users of IoT home security devices. The data collected during the pilot were subjected to tests of reliability and confirmatory factor analysis according to each construct and its antecedents. The authors decided to include all items in the pilot study in the primary data collection effort. In December 2017, data was collected through Cint from 229 U.S. consumers [42], using the same screening procedure used in the pilot. Demographics are shown in Table 4.1, and descriptive statistics for the measures are listed in Table 4.2. Based on an analysis of normal Q-Q plots for each variable, the data are normally distributed. Note that the sample skews heavily female. Both popular press and academic articles indicate that women are more likely than men to purchase IoT devices for safety (19% to 13%), while men are more likely to purchase IoT devices for energy savings [43,44]. Cannizzaro and colleagues [30] reported from their survey that while 57% of men owned smart home devices, 64% of women did. Given our interest here in privacy concerns in this study, it is interesting to note that the same authors reported that trust in privacy related to smart home devices was lower for women than for men.

Table 4.1 Sample demographics.

Gender	Education	Income
Female: 77.3%	Grad degree: 14.4%	<$50k: 55.5%
Male: 22.7%	College degree: 29.7%	$50k–$100k: 28.8%
	Some college: 27.5%	>$100k: 16.7%

https://doi.org/10.1371/journal.pone.0257601.t001

Table 4.2 Descriptive statistics (N = 229 for all variables).

	Minimum	Maximum	Mean	Standard Deviation
Intent	1.00	7.00	3.28	1.66
Fear	1.00	7.00	3.52	1.73
Attitude	1.00	7.00	5.11	1.37
Privacy Concerns	1.00	7.00	4.61	1.15
Subjective Norms	1.00	7.00	3.52	1.39

https://doi.org/10.1371/journal.pone.0257601.t002

Measurement Model

The measurement and structural models were both tested with AMOS 25, part of IBM's SPSS statistical package. We ran a confirmatory factor analysis with the five constructs in our model (Fig 4.2). The data fit the model well (Table 4.3). The reliability statistics for the scales were all at acceptable levels, and the average variance explained (AVE) values were all above 0.5 (Table 4.4). The square roots of the AVEs for each construct exceeded the largest correlation for that construct (Table 4.5). In summary, the measurement model demonstrated good convergent and discriminant validity.

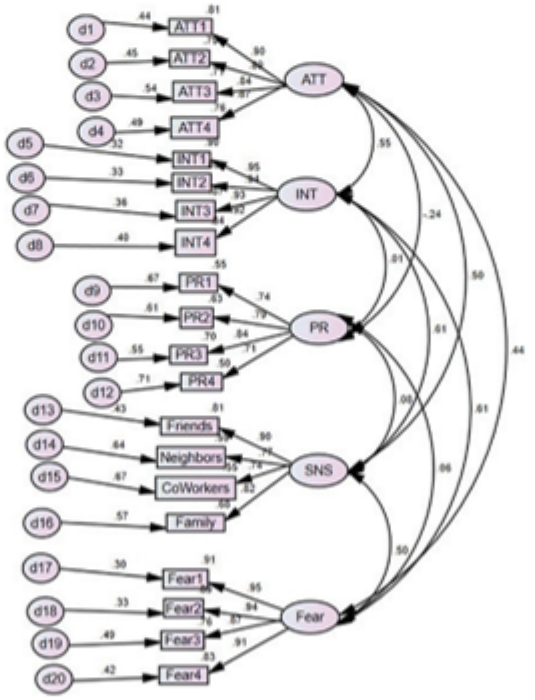


Fig 4.2 Confirmatory factor analysis results.

https://doi.org/10.1371/journal.pone.0257601.g002

Table 4.3 Measurement model fit statistics.

Criterion	Fit statistic
Chi-square	231.109
DF	160
Chi-square/DF	1.444
p	.000
CFI	.983
NFI	.946
RMSEA	.044

https://doi.org/10.1371/journal.pone.0257601.t003

Table 4.4 Measurement model convergent validity statistics (standardized regression weights).

Construct & item	λ	Cronbach's α	CR	AVE
Intention		.965	.966	.875
INT1	.948			
INT2	.942			
INT3	.932			
INT4	.919			
Attitude		.929	.930	.770
ATT1	.900			
ATT2	.892			
ATT3	.845			
ATT4	.872			
Privacy Concerns		.853	.854	.596
PR1	.744			
PR2	.793			
PR3	.836			
PR4	.707			
Fear		.956	.957	.846
Fear1	.953			
Fear2	.943			
Fear3	.872			
Fear4	.909			
Subjective Norms		.883	.884	.657
Friends	.901			
Neighbors	.767			
Co-workers	.741			
Family	.823			

https://doi.org/10.1371/journal.pone.0257601.t004

Table 4.5 Measurement model discriminant validity statistics.

	Intention	Attitude	Privacy	Fear	Subjective Norms
Intention	**.935**				
Attitude	.552	**.877**			
Privacy	.011	-.239	**.772**		
Fear	.611	.438	.062	**.920**	
Subjective Norms	.610	.499	.076	.496	**.810**

https://doi.org/10.1371/journal.pone.0257601.t005

Following [45], we took procedural remedies to control for common method variance (CMV). Specifically, we (1) kept the survey anonymous to reduce respondents' evaluation apprehension and to reduce the odds of respondents providing socially desirable, lenient, or acquiescent answers; (2) randomized the question order to control for priming effects and other biases related to question context or item embeddedness; and (3) as noted, improved scale items through the rigorous scale development process. With the survey data collected, we performed empirical tests to assess the extent of bias that may have resulted by CMV [46]. Following Lindell and Whitney [47], we partialed out a marker variable (a theoretically unrelated construct with three items: "I prefer blue to other colors," "I like blue color," and "I like blue clothes") in our model and examined the correlations between this marker and all latent variables. Common method bias exists if the correlation between any of the latent variables and the marker is greater than 0.3, which is not the case in this study, as the highest correlation we observed was 0.109. Following Podsakoff et al. (2003), we compared the R^2 changes of all endogenous variables in our model, before and after partialing out the above marker variable. Common method bias exists when there are significant changes of R^2 values; we found the highest R^2 value change in our data set to be 0.004. Common method bias is, therefore, not a threat to our study.

In addition, we tested for the possible presence of multicollinearity. Scatter plots showed that the data were normally distributed. Variance inflation factors for privacy concerns, fear, and subjective norms were all between 1.0 and 1.15, indicating multicollinearity was not a problem.

Test of TRA Structural Model

The test of the structural model yielded a $\chi 2\ (185) = 386.8$ ($\rho < 0.000$). Other indicators also pointed to a good fit of the data to the model: CFI = 0.951; NFI = 0.911; RMSEA = 0.069. Privacy concerns showed a negative and significant relationship with attitude toward purchasing IoT home security devices ($\beta = -0.26$, $\rho = 0.000$). Fear showed a positive and significant relationship with attitude ($\beta = 0.46$, $\rho = 0.000$). Attitude ($\beta = 0.40$, $\rho = 0.000$) and subjective norms ($\beta = 0.44$, $\rho = 0.000$) both showed a positive and significant relationship with intent to purchase. We tested the relationships between each control variable and the measure of intent, and none of these relationships was statistically significant except experience with burglary ($\beta = 0.16$, $\rho = 0.005$), such that those with personal experience with burglary had more intent to purchase. The model explained 38% of the variance in intent to purchase an IoT home security device and 28% of the variance in attitude (adjusted R^2). All four hypotheses were supported. The evaluated TRA model is shown in Fig 4.3.

Test of Mediation Structural Model

Including the direct paths from privacy concerns and fear to intent to purchase allow us to test the extent to which attitudes mediate the effect of beliefs on intent (Fig 4.4). The test of the mediation structural model yielded a $\chi 2\ (183) = 354.4$ ($p < 0.000$). The model was a good fit to the data (CFI = .959; NFI = .919; RMSEA = .064). Once again, attitude ($\beta = 0.29$, $p = 0.000$) and subjective norms ($\beta = 0.33$, $p = 0.000$) both

demonstrated positive and statistically significant relationships with intent. The effect on attitude of privacy concerns ($\beta = -0.26$, $p = 0.000$) was negative and statistically significant, and the effects of fear ($\beta = 0.45$, $p = 0.000$) were positive and statistically significant. These results are almost identical to the effects in the TRA model. However, whereas fear also had a positive and statistically significant direct effect on intent ($\beta = 0.38$, $p = 0.000$), privacy concerns did not ($\beta = 0.01$, $p = 0.837$). Privacy concerns effects on intent were completely mediated by attitudes. Experience with burglary also had an effect on intent ($\beta = 0.11$, $p = 0.036$), as was the case with the TRA model. The model explained 45% of the variance in intent to purchase an IoT security device and 27% of the variance in attitude toward purchasing (adjusted R^2).

Fig 4.3 Evaluated TRA structural model.

https://doi.org/10.1371/journal.pone.0257601.g003

Fig 4.4 Evaluated mediated structural model.

https://doi.org/10.1371/journal.pone.0257601.g004

Discussion

Using the TRA framework, we demonstrated that the intent to purchase IoT home security devices was about equally influenced by attitude toward purchasing and by the views of important others. Those important others included friends, family, neighbors and co-workers. If the respondent believed these people thought purchasing such devices was important, then the respondent had a stronger intent to purchase them. Attitude toward purchasing was based on beliefs about fear, related to home burglary, and beliefs about threats to privacy associated with security devices. Beliefs about privacy negatively affected attitude, but not as much as fear positively affected it. Despite concerns about an IoT home security device being associated with the loss of privacy, respondents were still driven by fear of home burglary to intend to purchase these devices. Concerns about privacy risks are not strong enough to undo the perceived benefits of IoT home security devices. To many consumers, the exchange of private information for security is worth the risk. In addition, experience with home burglary, whether of their own home or of others', increased respondents' purchasing intent.

Although the results supported all of our hypotheses and may seem obvious, they point to a more nuanced situation than might otherwise be expected. While fear was a major driver of attitude, and attitude drove intent, beliefs about the expectations of important others was just as important as attitude. This may be because the focus was on intent to purchase a very specific type of home security device–IoT devices–rather than home security devices generally. The views of important others may not be as important for devices that do not sit on one's home network but rather are installed and monitored by third parties. While we found that privacy concerns were also important to attitude, their importance was no doubt enhanced by the public perception of IoT devices as "leaky" sources of private data. Dramatic examples of voyeur access to children via IoT devices, such as the Mississippi Santa Claus case mentioned earlier, may also have played a role in enhancing the primacy of privacy concerns. We did not hypothesize about the potential role of control variables, and only experience with burglary was statistically significant. This finding showed that intent to purchase was enhanced by direct or indirect experience with home burglary.

Our mediation analysis showed that privacy concerns were completely mediated by attitudes towards purchasing IoT home security devices. Although we did not have specific expectations about this relationship, our findings are intuitive. As we have discussed, IoT home security devices are perceived by the public as being sources of leaked personal data or as facilitators of evasions of privacy. If a consumer believed an IoT security device was not secure, their attitudes towards purchase would be weakened, and the consumer's intent to purchase the devices would be lessened. Fear, on the other hand, proved to be a more general belief, mediated in part by attitudes, but also exerting a strong direct influence on intent to purchase. Fear could motivate a consumer to purchase an IoT security device, even if their attitudes toward purchasing such devices were not particularly strong. Privacy concerns, then, are dependent on the security technology in particular, while fear is less so.

From a theoretical perspective, TRA was shown once again to be a reliable, straightforward way to think about and study intent to engage in a particular behavior, this one being intent to purchase IoT home security devices. Although our central focus was on comparing the respective roles of fear and privacy concerns in affecting attitude, through relying on TRA, we also uncovered the importance of the perceived attitudes of others in formulating intent. By violating TRA's prescriptions about the central role played by attitudes, and conducting a mediation analysis, we were able to learn more about the relationships between beliefs and intent. Future research that relies on TRA might consider including mediation analyses of some sort.

From a practical perspective, it seems marketers of IoT home security devices are wise to focus on fear for increasing sales. They are also wise to downplay or not mention at all any IoT capabilities that might be associated with compromised privacy. Our findings indicate marketers may want to enhance fear by playing on consumer experience with home burglaries. They may also want to invoke how important others would

feel about purchasing IoT home security devices. A winning spiel might sound like "Home burglaries are more common than you think–they happen often and even in your neighborhood. Protect your home and family with XYZ IoT home security devices. Your mom wants you to keep her grandkids safe."

As is the case with any study, ours had limitations. The subject pool was limited to American consumers. The sample size satisfied the requirement of five observations per model parameter, but a larger sample size would permit the exploration of other beliefs and their effects on attitude. Using the Theory of Planned Behavior rather than the Theory of Reasoned Action would have necessitated collecting data on perceived behavioral control. However, the target behavior was purchasing IoT home security devices, which is a fairly straightforward behavior, both online and in person, so perceived behavioral control would probably have not played a major role. There may also be differences in attitudes toward purchasing security devices that require third-party monitoring, as opposed to the types of devices we were studying. Each of these limitations could be explored in future research, with a non-American sample, a larger sample, a different theoretical framework, and/or a different device to be purchased.

Conclusion

The global market for the smart home industry is projected to be worth over $150 billion USD by 2023, and a big component of that market is security devices, such as the Ring doorbell. The global market for smart home security devices alone is expected to exceed $5 billion USD by 2025. Fear is often used as a way to sell such devices, and we found that fear has a strong impact on consumer attitudes toward purchasing smart home security devices. Not surprisingly, we also found that security concerns about these devices also affected attitudes towards purchasing them. Attitude greatly influenced intent to buy, but what important others thought, and past experience with burglary, also affected intent. Attitude, subjective norms, and experience together explained 38% of the variance in intent. Researchers and marketers should take all three factors into account when exploring the major drivers behind purchasing IoT home security devices.

Supporting Information

S1 Appendix. Survey instrument.
(DOCX)

Author Contributions

Conceptualization: Joey F. George.
Data curation: Rui Chen.
Formal analysis: Rui Chen, Lingyao Yuan.
Investigation: Joey F. George.
Methodology: Joey F. George, Lingyao Yuan.
Project administration: Joey F. George.
Supervision: Joey F. George.
Validation: Rui Chen.
Writing – original draft: Joey F. George, Rui Chen, Lingyao Yuan.
Writing – review & editing: Joey F. George, Rui Chen, Lingyao Yuan.

References

1. Martin, C. The Internet Of Things vs. IoT hype. MediaPost, 2020. Available from https:// www.mediapost.com/publications/article/348679/the-internet-of-things-vs-iot-hype.html. Accessed 6/22/2020.

2. Statista, https://www.statista.com/outlook/279/109/smart-home/united-states 2018. Accessed 5/14/2020.
3. ADT. https://www.adt.com/home-is-where-the-smart-is. 2019. Accessed 5/14/2020.
4. Statistica, https://www.statista.com/statistics/1056057/worldwide-smart-home-security-market-value/#statisticContainer, 2021. Accessed 5/21/2021.
5. Zucker A. Amazon isn't the only one shopping for smart home security devices–US sales up 121.7 percent. The Brief Blog: Consumer Electronics, 2019; https://www.rakutenintelligence. com/blog/2018/amazon-isnt-the-only-one-shopping-for-smart-home-security-devices-us-sales-up-121-percent. Accessed 6/17/2021.
6. Molla, R. Amazon Ring sales nearly tripled in December despite hacks. Vox. 2020. Available from: https://www.vox.com/recode/2020/1/21/21070402/amazon-ring-sales-jumpshot-data. Accessed 5/14/2020.
7. Safewise. Do I really need a security system? 2020. Available from: https://www.safewise.com/ home-security-faq/do-i-need-a-security-system/ accessed 5/14/2020.
8. Reolink Blog. https://reolink.com/why-you-need-a-home-security-system/ 2018. Accessed 5/14/2020.
9. Schwarz, J. Ring gets 'dinged' for its video doorbell privacy. The Hill. 2020. Available from: https://thehill.com/opinion/cybersecurity/485449-ring-gets-dinged-for-its-video-doorbell-privacy. Accessed 5/14/2020.
10. Burke, M. Hacker accessed 'Ring' camera inside little girl's room, her family says. NBC News. 2020. Available from: https://www.nbcnews.com/news/us-news/man-hacks-ring-camera-8-year-old-girl-s-bedroom-n1100586. Accessed 5/14/2020.
11. Fishbein M., and Ajzen I. Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research. 1975. Reading, MA: Addison-Wesley.
12. Fishbein M., Jaccard J., Davidson A. R., Ajzen I., and Loken B. Predicting and understanding family planning behaviors, in Ajzen I. & Fishbein M. (eds), Understanding Attitudes and Predicting Social Behavior. 1980. Englewood Cliffs, NJ: Prentice Hall.
13. Ajzen I. From intentions to actions: a theory of planned behavior, in Kuhl J. & Beckman J. (eds), Action Control. 1985. Berlin: Springer, 11–39.
14. Sheeran P. & Taylor S. Predicting intentions to use condoms: a meta-analysis and comparison of the theories of reasoned action and planned behavior. Journal of Applied Social Psychology 1999: 29(8): 1624–1675.
15. Ruan, X. A study on the factors influencing chinese consumers' purchase intention on Volkswagen automobiles—focusing on country-of-origin image and corporate image. 4th International Education, Economics, Social Science, Arts, Sports and Management Engineering Conference (IEESASM 2016). 2016.
16. Paul J., Modi A. & Patel J. Predicting green product consumption using theory of planned behavior and reasoned action. Journal of Retailing and Consumer Services 2016; 29: 123–134. https://doi.org/10.1016/j.jretconser.2015.11.006.
17. Mady T. What makes up intentions to purchase the pioneer? a theory of reasoned action approach in India and the USA. International Journal of Emerging Markets, 2017;13(5): 734–757.

18. Nguyen N., Nguyen H.V., Nguyen P.T., Tran V.T., Nguyen N.N., Nguyen T.M.N., Cao T.K. & Tran H.N. Some key factors affecting consumers' intentions to purchase functional foods: a case study of functional yogurts in Vietnam. Foods, 2017; 9: 24. https://doi.org/10.3390/foods9010024 PMID:31881653

19. Yu S. & Lee J. The effects of consumers' perceived values on intention to purchase upcycled products. Sustainability, 2019; 11: 1034; https://doi.org/10.3390/su11041034

20. Rodrigo P., Khan H. & Ekinci Y. The determinants of foreign product preference amongst elite consumers in an emerging market. Journal of Retailing and Consumer Services 2019; 46: 139–148. https://doi.org/10.1016/j.jretconser.2018.04.012.

21. Lee S.H.N. & Chow P-S. Investigating consumer attitudes and intentions toward online fashion renting retailing. Journal of Retailing and Consumer Services 2020; 52. https://doi.org/10.1016/ j .jretconser.2019.101892.

22. Patil, K. Retail adoption of Internet of Things: applying TAM model. 2016 International Conference on Computing, Analytics and Security Trends (CAST), 2016. College of Engineering Pune, India. Dec 19–21, 2016.

23. Guo L. & Bai X. A unified perspective on the factors influencing consumer acceptance of Internet Of Things technology. Asia Pacific Journal of Marketing and Logistics, 2014; 26(2): 211–231.

24. Hsu C.-L. & Lin J. C.-C. An empirical examination of consumer adoption of internet of things services: network externalities and concern for information privacy perspectives. Computers in Human Behavior, 2016; 62: 516–527.

25. Aldossari M.Q. & Sidorova A. Consumer acceptance of internet of things (IoT): smart home context. Journal of Computer Information Systems, 2018. https://doi.org/10.1080/08874417.2018 .1543000

26. Klobas J.E., McGill T. & Wang X. How perceived security risk affects intention to use smart home devices: a reasoned action explanation. Computers & Security 2019; 87. https://doi.org/10.1016/j .cose.2019.101571.

27. Pinochet L.H.C., Lopes E.L., Szulzon C.H.F. & Onusic L.M. The influence of the attributes of 'Internet Of Things' products on functional and emotional experiences of purchase intention. Innovation & Management Review 2018; 15(3): 303–320.

28. Emami-Naeini, P., Dixon, H., Agarwal, Y. & Cranor, L.F. Exploring how privacy and security factor into iot device purchase behavior, in CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK. ACM, Glasgow, UK, 12 pages. https://doi.org/10.1145/3290605.3300764.

29. Mani Z. & Chouk I. Consumer resistance to innovation in services: challenges and barriers in the Internet of Things era. Journal of Product Innovation Management 2018; 35(5): 780–807.

30. Cannizzaro S., Procter R., Ma S., and Maple C. Trust in the smart home: Findings from a nationally representative survey in the UK. PLoS ONE 2020, 15(5): e0231615. https://doi.org/10.1371/journ al.pone.0231615. Accessed 5/26/2021. PMID:32469883

31. Rogers R. W. A protection motivation theory of fear appeals and attitude change. Journal of Psychology 1975; 91(1): 93–114.

32. Boss S., Galletta D., Lowry P. B., Moody G. D., and Polak P. What do systems users have to fear? using fear appeals to engender threats and fear that motivate protective security behaviors. MIS Quarterly 2015; 39(4): 837–864.

33. George J. F. The theory of planned behavior and internet purchasing. Internet Research 2004; 14(3):198–212.

34. Jarvenpaa S. L., and Todd P. A. Is there a future for retailing on the Internet. Electronic Marketing and the Consumer 1997; 1(12): 139–154.

35. Swaminathan V., Lepkowska-White E., and Rao B. P. Browsers or buyers in cyberspace? an investigation of factors influencing electronic exchange. Journal of Computer-Mediated Communication 1999; 5(2): JCMC523.

36. Cooney, M. FTC: Best practices for businesses facing internet of things onslaught. Networkworld, Jan 27 2015. Available from: https://www.networkworld.com/article/2876312/ftc-best-practices-for-businesses-facing-internet-of-things-onslaught.html; Accessed 5/29/2020.

37. Bandara R., Fernando M. and Akter S. Privacy concerns in e-commerce: a taxonomy and a future research agenda. Electronic Markets, 2019. https://doi.org/10.1007/s12525-019-00375-6.

38. Fortes N. and Rita P. Privacy concerns and online purchasing behaviour: towards an integrated model. European Research on Management and Business Economics 2016. 22: 167–176.

39. Taylor S., and Todd P. A. Understanding information technology usage: a test of competing models. Information Systems Research 1995; 6(2): 144–176.

40. Milne S., Sheeran P., and Orbell S. Prediction and intervention in health-related behavior: a meta-analytic review of protection motivation theory. Journal of Applied Social Psychology 2000; 30(1): 106–143.

41. Xu H., Dinev T., Smith J., and Hart P. Information privacy concerns: linking individual perceptions with institutional privacy assurances. Journal of the Association for Information Systems 2011; 12(12): 798.

42. George, J. Internet-of-Things (IoT) home security purchasing data, December 2017. 2021. Iowa State University. Dataset. https://doi.org/10.25380/iastate.14036000.v1.

43. hippo.com. Stats on home protection. https://www.hippo.com/blog/smart-home-statistics. 2020. Accessed 5/21/2021.

44. Halliday, A. How age and gender affect interest in smart home technology. 2016. https://realtybiznews.com/how-age-and-gender-affect-interest-in-smart-home-technology/98732531/. Accessed 5/21/2021.

45. Podsakoff P. M., MacKenzie S. B., Lee J.-Y., and Podsakoff N. P. Common method biases in behavioral research: a critical review of the literature and recommended remedies. Journal of Applied Psychology 2003; 88(5): 879. https://doi.org/10.1037/0021-9010.88.5.879 PMID:14516251

46. Tehseen S., Ramayah T., and Sajilan S. Testing and controlling for common method variance: a review of available methods. Journal of Management Sciences 2017. 4(2): 142–168.

47. Lindell M. K., and Whitney D. J. Accounting for common method variance in cross-sectional research designs. Journal of Applied Psychology 2001; 86(1): 114. https://doi.org/10.1037/0021-9010.86.1.114 PMID:11302223

4.8 Case Study 2

RESEARCH ARTICLE

Piloting a digital campaign to promote awareness of the Louisiana TelePrEP program among sexual and gender minority young adults

Manuel A. Ocasio[1]*, M. Isabel Fernandez[2], Steven Cortese[3], Kathryn Kampa[4]

1 Department of Pediatrics, School of Medicine, Tulane University, New Orleans, Los Angeles, United States of America, 2 College of Osteopathic Medicine, Nova Southeastern University, Fort Lauderdale, Florida, United States of America, 3 Communify, New Orleans, Los Angeles, United States of America, 4 Department of International Health and Sustainable Development, School of Public Health and Tropical Medicine, New Orleans, Los Angeles, United States of America

* *mocasio@tulane.edu*

Abstract

Despite the proven efficacy of pre-exposure prophylaxis (PrEP) in preventing HIV acquisition, PrEP is underutilized by sexual and gender minority young adults in the 13 states in the United States "South." In this paper, we describe the process we used to develop a digital campaign to promote awareness and engagement of sexual and gender minority young adults in the Louisiana Department of Health's TelePrEP Program and provide campaign performance metrics. In Phase 1, we conducted formative research that informed campaign development and strategy. In total, 109 sexual and gender minority young adults completed a survey of PrEP constructs (e.g., facilitators, barriers). We also conducted three, sequential focus groups to iteratively generate, revise and refine the digital material. In collaboration with our strategic marketing partner (SMP), we developed 3 different ads and 1 video ad to promote on web and mobile-in app display, as well as Facebook, Instagram, and YouTube. Phase 2 focused on campaign implementation and evaluation (e.g., number of impressions, user activities on LA TelePrEP landing page). In the first few weeks of the campaign, data from tracking pixels indicated minimal activities on the landing page. We paused to revamp the campaign. Our SMP determined that a more young adult-focused landing page could bolster engagement. We created a new landing page and reran the campaign for 33 days. We saw substantially more user activities on the new landing page (n = 382) compared to the LA TelePrEP landing page (n = 185). Overall, we had 730,665 impressions and 475 link clicks. By collaborating with our SMP, we effectively and efficiently translated our community-engaged formative research into relevant and engaging digital content. This pilot study is one of the first to demonstrate the importance of using tracking pixels to monitor real-time user data to optimize performance of a digital PrEP campaign.

Introduction

Despite the advances of HIV prevention efforts in many areas of the United States, Louisiana and the other 13 states (Texas, Oklahoma, Arkansas, Mississippi, Alabama, Georgia, Florida, North and South Carolina, Tennessee, Kentucky, Virginia, West Virginia) referred to as the "South" in the HIV surveillance reports continue to be heavily impacted. The "South" accounted for 51% of the new HIV diagnoses in 2020 [1].

https://doi.org/10.1371/journal.pone.0290149 Used with CC 4.0 license

Furthermore, the "South" has the highest rate of HIV in rural areas and fewer people who are aware of their HIV status compared to other regions of the country. Sexual minority young adults (persons with non-heterosexual identity, attraction, and/or sexual contact) and gender minority young adults (persons whose gender identity or expression is different from their assigned sex at birth) [2] in the "South" are at particularly elevated risk relative to their cisgender heterosexual counterparts. In the "South," cisgender gay, bisexual and other men who have sex with men, particularly younger men, bear the brunt of the HIV epidemic [3]. Limited data also suggest that 14% of transgender women and 2% of transgender men are living with HIV, a majority of whom are in the "South" [1]. In Louisiana, adolescents and young adults 13 to 24 years of age comprised 21% of all new HIV cases diagnosed in 2019. Almost half of these new cases were concentrated in New Orleans and East Baton Rouge [4].

Given these sobering statistics, it is not surprising that approximately half of all jurisdictions in the Ending the HIV Epidemic initiative (EHI) are situated in the "South." The EHI is a U.S. Department of Health and Human Services initiative designed to reduce new HIV infections by scaling up HIV prevention and treatment strategies in the areas with the highest concentration of new HIV cases diagnosed [5]. To achieve the target of 90% reduction of new HIV infections by 2030 [5], prevention efforts must be ramped up in EHI jurisdictions so that those most at-risk can access and utilize effective HIV prevention methods.

Pre-exposure prophylaxis (PrEP) is a potent and efficacious strategy to prevent HIV acquisition. PrEP has been found to effectively reduce HIV transmission in a number of groups from diverse geographic areas including adult cisgender gay, bisexual and other cisgender men who have sex with men, people of transgender experience, heterosexual men and women, and serodiscordant couples [6]. The efficacy of PrEP, when taken consistently, has also been demonstrated among young cisgender gay, bisexual and other men who have sex with men [7, 8]. When taken as prescribed, PrEP can reduce the risk of getting HIV from sex by about 99% [9]. Yet, PrEP is underutilized particularly by racial and ethnic sexual and gender minority adolescents and young adults from EHI jurisdictions in the "South" [10] such as New Orleans and East Baton Rouge, Louisiana. For instance, in 2021, the PrEP-to-Need Ratio (PNR), a metric that reflects whether PrEP use appropriately reflects the need for HIV prevention, was lowest among young people aged 13 to 24 relative to all other age groups in Louisiana [11]. Stigma related to HIV and PrEP use have been consistently cited as barriers to PrEP uptake [12, 13]. A 2022 study of Black and Latinx sexual minority cisgender men and transgender women found that PrEP stigma was independently associated with current PrEP use; however, when researchers tested for the moderating effect of logistical barriers, PrEP stigma was only significant when logistical barriers to PrEP use were low [14]. Thus, programs that can reduce logistical barriers to PrEP access, while addressing related stigma among sexual and gender minority young adults are needed to increase PrEP use.

To increase PrEP access, the Louisiana Department of Health launched the Louisiana TelePrEP Program (LA TelePrEP) in 2018. This novel telemedicine program connects Louisiana residents 18 years of age or older to a provider who prescribes PrEP and to an e-navigator who facilitates enrollment in the program, lab testing, insurance access, and other ancillary services. Enrollment can be done remotely by completing the form accessible on the LA TelePrEP website [15]. Telemedicine approaches can reduce stigma which is often experienced in face-to-face clinical settings and eliminate commonly cited logistical barriers, such as transportation [16]. Unfortunately, relatively few sexual and gender minority young adults have enrolled in the program despite outreach and promotional efforts launched at local colleges, distribution of flyers at community-based organizations (CBOs) and clinics, and posts through the health department's social media profile. Clearly, the department's efforts were not reaching young adults who are sexual and gender minority; alternative strategies for increasing the sexual and gender minority young adult's participation in the TelePrEP program were sorely needed.

https://doi.org/10.1371/journal.pone.0290149

In 2019, the Adolescent Trials Network for HIV/AIDS Interventions (ATN) published a call for applications with a short project period to advance the goals of the EHI targeting adolescents and young adults. Partnering with health departments and/or community-based organizations (CBOs) was a cornerstone of the call given the pivotal role these entities play in the fight against HIV. In response to this call, ATN affiliated researchers from two universities in the South partnered with the Louisiana Department of Health to develop and pilot a digital campaign to increase awareness of the LA TelePrEP program among young adults who are sexual and gender minority 18 to 29 years of age.

Digital technologies offer great promise for increasing awareness of the LA TelePrEP program among sexual and gender minority young adults. Ninety percent of young people regularly use digital technologies (e.g. social media) to gain and share information, post pictures, conduct financial transactions, stay current and build social connections [17, 18]. In addition to being accessible, familiar and user-friendly, online tools available on many social media sites enable the direct purchase of banner ads [19], at reasonable and affordable price points. These ads could be used for a variety of purposes including providing information about programs and services, directing users to different landing pages, and prompting users to take specific actions (e.g., complete a screener). Furthermore, many digital platforms such as Google, Facebook and social media apps collect data on users' interest and online behavior. These data can be used to target specific subpopulations, such as young adults who are sexual and gender minority in Louisiana, efficiently and cost-effectively as well as direct ads to individuals who are most likely to engage with them [19, 20]. It is not surprising that there has been a dramatic increase in the use of technology for HIV prevention research and practice.

Given the challenges of engaging hard-to-reach populations such as young adults who are sexual and gender minorities, significant attention has been focused on social media and dating app-based advertisements to increase awareness of on-going research studies and enhance recruitment [21–23]. These studies provide strong evidence that digital ads can be used to encourage sexual and gender minority young adults to engage with research studies by completing the screening questionnaires. Many of these studies use impressions, clicks, click through rate (CTR) and cost per click as evaluation metrics. Until recently, relatively little attention has been given to comparing across different social media apps and platforms. Zlotorzynska, Bauermeister, Golinkoff et al [23] used engagement, screening and cost metrics to compare the performance of ads placed on Facebook, Instagram, Snapchat, Twitter and Grindr to recruit young adults who are sexual and gender minorities across the United States. They found considerable differences in performance across the platforms. For instance, Instagram proved most effective for reaching young, eligible racial and ethnic minorities. The lowest cost per eligible youth were ads placed on Facebook properties. Furthermore, there were regional differences in cost per eligible contact.

Dehlin, Stillwagon, Pickett et al [24] reported on a grassroots effort to develop and launch a multi-media campaign that included billboards, posters and other print ads as well as digital media outlets to promote PrEP awareness and use among sexually-active people in Chicago. Advertisements directed users to the campaign website and a hotline for a warm handoff to PrEP services. They evaluated the performance of both the print and the social media ads during a four-month period. Across social media platforms and display advertising, their ads were viewed 40 million times and the campaign website averaged 182 clicks per day. They estimated that 83 people who saw the ads contacted the hotline and 18% started PrEP. The successful use of digital media ads for awareness and engagement, support developing a targeted digital campaign to increase awareness of the TelePreP program among young people who are sexual and gender minorities in Louisiana. In this paper, we describe the process we used to develop the digital campaign in hopes that others could use our experiences to guide their campaign development efforts. Since the goals of our modest campaign were to increase awareness of the TelePrEP program and engagement with the campaign landing page, we provide campaign performance metrics to illustrate how the campaign impacted user awareness and engagement.

https://doi.org/10.1371/journal.pone.0290149

Methods

In Phase 1, the formative research phase, we conducted a community survey and sequential focus groups that guided and informed the development of the campaign material and strategy. We partnered with a locally based LGBTQ+ strategic marketing agency. We also designed the protocols for launching the campaign and defined the metrics we would use to evaluate its reach. The materials, protocols and metrics were developed iteratively with extensive input from our partners and community members. In tandem, we iteratively developed a chatbot that would support the e-navigator and disseminate information about PrEP and the TelePrEP program. We describe our chatbot development protocol in a separate manuscript [25]. Phase 2 focused on implementation and evaluation of the digital campaign and launch of the chatbot. The project period was 20 months.

The study was approved by the Tulane University Institutional Review Board (2021–107). Only IRB-approved authors and study staff had access to identifiable participant information that was collected during Phase 1 formative research. These data will be destroyed once study dissemination activities are completed.

Phase 1

Community survey: We developed and programmed a brief Qualtrics survey to identify key indicators that capture constructs related to PrEP uptake (experiences, intentions, beliefs, facilitators, barriers) in sexual and gender minority young adults. The majority of the survey items were drawn from the published literature and reviewed by our Youth Advisory Board, which consisted of 8 young adult members (88% ethnoracial minority, 100% sexual or gender minority). We used both passive and active recruitment methods. Active recruitment included directly approaching youth at community events, bars and clubs frequented by sexual and gender minority young adults and inviting prior study participants who had consented to be contacted for future studies. Participants were invited to refer others to the study. Additionally, Louisiana Department of Health and sexual and gender minority-serving community-based organization (CBO) staff shared recruitment materials with clients and encouraged them to participate in the study. For passive recruitment, we posted flyers at community venues frequented by sexual and gender minority young adults and paid advertisements on Instagram, Facebook, and Jack'd, a popular gay dating app. We also used a study-specific profile to disseminate the survey on social media sites. Young adults who completed the survey received $10 and an invitation to participate in the focus group component. Participants were compensated $5 for each person they referred who completed the study. All compensation was distributed via CashApp. We implemented multiple strategies to identify fraudulent entries, such as survey bot responses and duplicate entries. First, we created separate links for the screener, community survey and payment form to minimize the ability for a bot to complete all forms. Second, we programmed the survey to limit one response per IP address. Third, we used timestamps to determine how close in time consecutive surveys were completed and the length of time for survey completion to differentiate rapid responses by a bot from a human being.

From June, 2021 to March, 2022 we enrolled 109 young adults who: 1) self-identified as sexual and/or gender minority; 2) were 18 to 29 years of age; 3) reported condomless sex in the last 6 months; 4) lived in Louisiana; 5) and were HIV negative or unknown HIV status. Thirteen participants were recruited through referrals. On average, participants were 22 years of age; 85% came from the Greater New Orleans area and 7% came from East Baton Rouge. Sixty-two percent identified as cisgender men, 54% as gay and 26% as gender diverse (e.g., non-binary, genderqueer); and almost half (47%) were Black or African American. Although 89% of the sample had heard of PrEP, a majority had never taken PrEP (63%). Participants reported that the main facilitators of using PrEP were that it protects against HIV (55%), is easier to use than other methods (32%), and does not interrupt sex (31%). In contrast, the main barriers identified were cost (31%), having to take a daily pill (53%), and the potential for harmful side-effects (38%). Awareness of the TelePrEP program was

https://doi.org/10.1371/journal.pone.0290149

low; only 25% of the sample knew about the program. A majority of participants reported using Instagram (86%) and 40% reported using Instagram more than any other social media apps.

Focus groups: Guided by the survey results, we developed a focus group discussion manual that tapped the following areas: 1) social media use (likes, dislikes); 2) perceptions of HIV-related messaging on social media; and 3) recommendations for promoting TelePrEP on social media. We convened three focus groups on HIPAA-compliant Zoom that were conducted from April to July, 2022. Each focus group was comprised of 4–6 participants who had completed the on-line survey and was audio-recorded with permission. Participants received $40 as compensation via CashApp. All participants provided written consent electronically.

Focus groups were led by two trained facilitators. Participants viewed sample digital material and used a five-point scale (1 not at all 5 extremely) to anonymously rate the content of each sample as being: 1) engaging; and 2) convincing. After reviewing the responses for each component, participants discussed their ratings, what guided their decisions, and provided recommendations for improvement. We were especially interested in understanding participants' thoughts on visual (imagery, colors, font) and textual (tone, language, clarity) aspects of the content. Facilitators independently took notes during focus groups and reviewed recordings to modify notes, as needed. The focus groups followed the same procedures, although the sample digital content presented differed.

The content of the focus groups was developed in collaboration with our strategic marketing partner (SMP) and followed an iterative process to generate, revise and refine the digital material. For the first focus group, we presented digital material from existing PrEP-related promotional materials and selected the most salient PrEP related messages based on the survey result. We also included existing LA TelePrEP advertisements. The results of the first focus group were used to create the digital content for the second focus group and the results of the second focus group generated the content for the third focus group. The materials were again refined prior to being presented to the Louisiana Department of Health Material Review Committee and Bureau of Media and Communications for final review and approval. This process yielded the three digital advertisements (and banner ads) used in the campaign.

We identified consistent themes across the 3 focus groups that informed the final digital campaign creative and strategy. In terms of content, participants recommended that the materials include factual information about PrEP's efficacy and having an imprimatur from a known and credible organization such as the health department. Participants recommended highlighting aspects of the TelePrEP Program that would resonate with sexual and gender minority young adults such as privacy and home delivery. In terms of visuals, participants strongly preferred ads with more than one racially diverse, overtly queer young adult over those with infographics or text only. They encouraged us to use attention grabbing bright colors, motion and music and to advertise on social media. We considered these recommendations as we developed each iteration of the digital materials and the digital campaign strategy. In total, we developed 3 ads (Figs 1–3).

Digital campaign strategy: Armed with the results from the community survey (salience of PrEP efficacy, low awareness of TelePrEP program, social media preferences) and focus groups (bright colors, overt queerness, privacy, home delivery), we worked collaboratively with our SMP to define the goal of the digital campaign and develop a marketing strategy that would appeal to sexual and gender minority young adults. The primary goal of our digital campaign was to raise awareness of the TelePrEP program. We arrived at this decision because almost three-fourths of survey participants had not heard about the TelePrEP program and few sexual and gender minority young adults had enrolled in the TelePrEP program. We delineated the key performance indicators (KPI) for awareness as: 1) exposure measured through number of impressions; and 2) interactions measured as number of clicks and click-through rate (CTR). Even though the short project period limited the duration of the digital campaign, we set a secondary, exploratory goal of engagement with the campaign website to inform future efforts. To track engagement, we used two metrics: 1) the number

https://doi.org/10.1371/journal.pone.0290149

of visits to the landing page (the Louisiana Health Hub's TelePrEP Program landing page) resulting from exposure to our ads; and 2) user activities on the landing page. We placed tracking pixels on the LA TelePrEP landing page to log number of visits to the landing page, which advertising strategy led users to the landing page, and track user activities on the landing page. In addition to these performance indicators, we also tracked cost efficiency of each digital strategy (i.e., cost per click). See Table 4.6 for a description of metrics and key terminology.

Table 4.6 Descriptions of key terms.

Term	Definition
General Terms	
Tracking Pixel	Code embedded in a website (e.g., landing page) to track actions users take on a website.
Call to Action	Text or visual element that prompts a desired action.
Impressions	Number of times an advertisement was viewed (not unique users. One user can contribute to impressions more than once)
Link Clicks	Number of times a person clicked on the call to action button that leads to the website (e.g., landing page).
Click-Through Rate	Percentage of users who click on an advertisement (to enlarge, call to action to the number of impressions.
Link Click-Through Rate	Percentage of users who clicked on link to website (e.g., landing page) to the number of total number of impressions
Cost Per Click	Cost allocated to digital strategy (e.g., display) divided by number of link clicks.
Cost per Impression	Cost allocated to digital strategy (e.g., display) divided by number of impressions.
Site Activity	Prespecified actions that are tracked on a website (e.g., landing page).
Display Advertising Terms	
Contextual Targeting	Users identified based on presence of a keyword on a website they are visiting for future advertising.
Search Targeting	Users identified based on use of keywords in a search engine.
Mobile In-App Advertising	Ads displayed in mobile apps.
Extended Audience	Campaign feature that uses machine learning to identify similar users from the keyword list.
Social Media Advertising Terms	
Awareness Objective	Objective to maximize reach to users who are most likely to remember ad.
Traffic Objective	Objective to drive traffic to the website (e.g., landing page).
YouTube Terms	
Video Views	Number of users that viewed a video (equivalent to impressions).
Video Completion Rate	Percentage of users who viewed a video to the end.
Topic	Targeting method to show users video based on type of viewing content.
Audience	Targeting method to show videos based on user characteristics (interests, demographics)

https://doi.org/10.1371/journal.pone.0290149.t001

We decided to use display ads (web and mobile in-app) as well as ads posted on Facebook and Instagram since 86% of survey participants used Instagram and 66% used Facebook. To direct the placement and management of our display ads, we used the Google Ads advertising platform which provides a streamlined process for setting up a targeted digital campaign via search, contextual targeting and audience expansion algorithms (Google support) guided by a keyword list. For ads on Facebook and Instagram, we used Meta Ads Manager and selected awareness and traffic as our goals. Our target audience settings were set to target 18 to 24 year olds in zip codes with the largest population density in Louisiana [26]. We used Meta's algorithm, rather than manual selection, to optimize ad placement. Thus, as a campaign runs, Meta will place more ads on the platform that is performing better.

Phase 2

In September of 2022, we launched the digital campaign. The SMP monitored the campaign's performance and provided weekly updates. Three weeks into the campaign, KPI data showed a high number of impressions (exposure) and large numbers of link clicks (interactions). However, the tracking pixel data demonstrated minimal engagement with the campaign's landing page. Though engagement activities were exploratory, we were concerned that there was such dramatic inconsistency between ad engagement and website use. After 28 days, we paused the campaign to identify potential reasons for the lack of engagement and develop viable solutions. First, we evaluated the characteristics of the landing page itself. In contrast to the bright colors (bright orange, pink, turquoise) and youthful font of our ads, the TelePreP Program's landing page used more subdued colors (muted green and blue) and a more generic font that matched those used by the Louisiana Department of Health. Thus, the stark visual contrast between our ads and those on the landing page may have turned off users. Second, we examined the content. Rather than a direct and focused "call to action," the landing page offered users multiple "calls to action," which may have led users to feel overwhelmed, distracted and unwilling to select an option. For example, the landing page had comprehensive information on the program as well as provided options to schedule an appointment, text the e-navigator, and read FAQs. Third, we explored the feasibility of developing video content as recommended by focus group participants.

Based on this evaluation, we paused the campaign for 33 days and revised the digital strategy. We created a campaign-specific landing page that mirrored the colors and font used in our digital ads. The only call to action was directing users to text an SMS chatbot for more information [25]. Because we lacked sufficient resources and time to create a storyboard for the video and film its content, our SMP created a motion graphic video maintaining the campaign aesthetic and incorporating suggestions from our focus groups. We added YouTube as a display medium for our video, because of its popularity among young people [27]. The video ad that ran on YouTube can be viewed at: https://youtu.be/1FMKpB2m5TQ. The revised campaign ran for 40 days.

Results

Over the course of the campaign, we had more than 730,000 impressions and 475 link clicks. Table 4.7 shows data comparing KPI and cost efficiency across our digital advertising strategies. Facebook was the most efficient in terms of CTR (0.14%) and combined with Instagram, had the lowest cost per click ($8.62). Notably, Facebook outperformed Instagram, which was the most popular app among our study participants, in number of impressions and CTR. Web display advertising had the highest number of impressions (n = 458,172) and yielded the most link clicks overall (n = 278). Though YouTube was the least cost effective and had the fewest number of link clicks, there were 39,364 complete video views for a 66.6% video completion rate.

https://doi.org/10.1371/journal.pone.0290149

Table 4.7 Key performance indicators and cost efficiency across strategies.

Strategy	Impressions/Views	Link Clicks	CTR	Cost Per Click
Facebook	**108,105**	**149**	**0.14%**	**8.62**[a]
Instagram	**40,407**	**11**	**0.03%**	
Web Display	**458,172**	**278**	**0.06%**	**$8.89**
Display In-App	**64,859**	**22**	**0.03%**	**$22.73**
YouTube	**59,122**	**15**	**0.03%**	**$200.00**

[a]Cost per click could not be separated by platform.

https://doi.org/10.1371/journal.pone.0290149.t002

Table 4.8 summarizes the data generated by the tracking pixels. Over the course of the campaign, we tracked a total of 567 engagement activities across both landing pages. There was more than double the number of pixels fired on the new campaign-specific landing page when compared to the TelePrEP landing page even though the duration of the campaign periods were comparable. Beyond opening the website and allowing it to load, there was more engagement with the new landing page (viewing and clicking on "Learn More") relative to the TelePrEP landing page. Other calls to action that we tracked (e.g., scheduling a consultation and calling or texting the Program) had zero engagement by users and thus are not shown on the table.

Table 4.8 User activities on the TelePrEP landing page and new campaign-specific landing page.

Website Activity	Description	Total Activities
LA TelePrEP Landing Page		185
Landing Page	A user lands on the website and it loads.	165
PrEP Directory	A user opens the PrEP Directory page.	5
Learn More (Click)	A user clicks on the "LEARN MORE" button.	14
Click Here to Enroll (Click)	A user clicks on the "CLICK HERE TO ENROLL" button.	1
New Landing Page		382
Landing Page	A user lands on the website and it loads.	215
Learn More (Visibility)	A user scrolls down and the "LEARN MORE" button is visible.	104
Learn More (Click)	A user clicks on the "LEARN MORE" button.	42
Text Our PrEPBot (Click)	A user on a mobile device clicks on "TEXT OUR PREPBOT."	21

https://doi.org/10.1371/journal.pone.0290149.t003

Table 4.9 compares the performance of the ads. While KPIs for ads on web display was comparable (all 0.12% CTR), CTR varied in social media advertising. Overall, Facebook had higher CTR compared to Instagram and web display, specifically Versions A (CTR = 0.24%) and B (CTR = 0.23%). Version A delivered the most web activity (n = 191) on web display, while Version B delivered the most web activity on social media platforms (n = 36).

https://doi.org/10.1371/journal.pone.0290149

Table 4.9 Metrics for key performance indicators for different ad versions across social media platforms and display advertising.

	Social Media							Display			
	Facebook			Instagram			All Social Media[a]				
Ad Version	Impressions	Link Clicks	CTR	Impressions	Link Clicks	CTR	Any Web Activity	Impressions	Link Clicks	CTR	Any Web Activity
Version A	11,388	27	0.24%	4315	3	0.07%	11	152,590	176	0.12%	191
Version B	42,124	95	0.23%	16459	5	0.03%	36	151,479	189	0.12%	153
Version C	54,593	27	0.05%	4315	3	0.07%	2	151,269	187	0.12%	118

[a]Web activities could not be disaggregated by social media app.

https://doi.org/10.1371/journal.pone.0290149.t004

Discussion

The goal of our study was to develop and pilot a digital media campaign to increase awareness and engagement of sexual and gender minority young adults in the Louisiana TelePrEP program. Our study highlights the benefits of engaging community members and digital marketing experts in all phases of the campaign development and implementation process. In a rapidly evolving digital environment with changing trends and constant barrage of information, advancing digital-based HIV prevention efforts for sexual and gender minority young adults requires technical knowledge and expertise that few scientists possess. Researchers must expand their collaboration beyond traditional scientific disciplines to the business and technology sectors and partner with digital marketing experts such as our SMP. By collaborating with our SMP, we effectively and efficiently translated our community-engaged formative research into relevant and engaging digital content. Furthermore, by using tracking pixels to monitor campaign activities in real time, our SMP detected a problem with user engagement within 3 weeks of the launch and guided us to make the necessary adjustments to improve engagement, rapidly and effectively.

The strategy of using tracking pixels to monitor user behavior on the landing page in response to our ads is an important contribution suggested by our SMP that has not been frequently used in HIV prevention campaigns for sexual and gender minority young adults. Tracking pixels are a powerful tool that can be used to directly compare KPIs across digital platforms and different forms of advertising to evaluate progress towards campaign goals in real time. In our pilot, using tracking pixels to monitor campaign performance yielded the real-time data needed to refine our campaign strategy. Having performance data early proved particularly useful in our situation given our short project period where time was of the essence and our resources were limited.

Despite the benefits, tracking pixels do not provide information on the personal characteristics of the users; they only monitor user activity. Thus, we could not directly ascertain that all users tracked were sexual and gender minority young adults. Our digital campaign focused on increasing awareness of the TelePrEP program and user engagement with the landing page and not on recruiting eligible individuals to complete on-line surveys or participate in intervention trials. When social media is used as a recruitment strategy, the data to ascertain the demographic profiles of users is available from completed screeners, surveys or other measurement instruments [23, 28, 29]. To report with certainty that all of our users were sexual and gender minority young adults, we would have had to add a brief survey as a call to action on the landing page which was beyond the scope of our efforts. Because the content (e.g., the images, the poses, the captions) and placement/targeting of the ads were designed to appeal to sexual and gender minority young adults, it is likely that, at minimum, a proportion of our users were sexual and gender minority young adults. Other researchers' success in reaching and enrolling young sexual and gender minorities into studies using similar targeting approaches to ours [23], supports this perspective.

Our experience with tracking pixels has implications that extend beyond digital campaigns. Health promotion interventions and programs are increasingly using digital technologies [30, 31]. For instance, apps are being used as delivery vehicles [32], users are being referred to websites for information and resources [33]; and services are being advertised through social media. Embedding tracking pixels into health promotion interventions and programs holds tremendous promise for helping researchers and practitioners gain important data to make rapid adjustments and improve performance. However, this area is not without controversy as data privacy has been a primary concern in digital advertising. Industry and government have been gradually setting policies and regulations to limit tracking specific individuals. For example, Apple has instituted opt-in permissions for third-party apps to access user data [34]. In our case, our SMP did not collect identifiable information; we only collected data pertaining to user engagement on the landing page. Furthermore, in keeping with the increasing trend for respecting user privacy, our

https://doi.org/10.1371/journal.pone.0290149

partners maintain strict requirements with marketers to ensure that they will not use any user data without user consent. This reinforces the importance of collaborating with partners that have strong policies to protect against misuse of data.

In addition to monitoring user activities, it is important to examine the performance of the digital content per se and the placement of the content. Monitoring these parameters can yield real-time information to refine campaigns such as revising or removing lower performing content in favor of the better performers or shifting the placement of the ads. We saw no difference in CTR among the three versions of the digital content when they appeared as display ads on websites or apps. Yet in social media, Versions A and B outperformed Version C.

The performance of the YouTube video is worth noting even though the results should be interpreted with caution because of the limited time that the video was shown. Industry standards vary in terms of what constitutes a reasonable video completion rate, but some consider 15% to be acceptable [35] and 70–80% as good [36]. Our 66.6% rate falls on the higher end of this spectrum, suggesting that YouTube is promising for disseminating information. In terms of ad placement on social media platforms, we were surprised to find that Facebook yielded better KPIs for campaign awareness than Instagram. Results from our community survey and nationally-representative surveys indicate higher Instagram use among young adults compared to other social media apps [37]. Furthermore, social media marketing experts recommend Instagram over Facebook for effective brand awareness among younger populations [38]. However, because our campaign period was brief, we cannot definitively claim Facebook as a more effective platform for promoting awareness of TelePrEP programs among sexual and gender minority young adults. If we had had a longer project period, we would have enhanced our presence on social media and YouTube and made adjustments on placement for the ads. Additionally, we would have reduced the budget for display ads, particularly for mobile in-app display. It is likely that these revisions would have yielded better performance at the same or reduced cost. Having real-time data during implementation can be extraordinarily helpful to improve campaign performance as our findings have shown.

Despite the novel findings, there were some limitations to our study. As with other digital campaigns, we did not have a way to ascertain the exact demographic profile of users who viewed and engaged with ads and the landing pages. Thus, we cannot be certain that our users were actually young sexual and gender minorities even though we designed the campaign to target them. Furthermore, the short timeframe and small budget significantly limited our ability to more fully evaluate the performance of our digital campaign. Use of both print and social media ads over a longer period would have enhanced the performance reaching more users. Research has shown that multiple exposures to ads via different outlets is a key element for effective health communication campaigns [39]. With more resources and time, we could have maximized reach by targeting other social media apps, as well as popular gay dating apps that have successfully engaged youth in HIV prevention research [23, 40]. Lastly, we would have worked with the Louisiana Department of Health to design a TelePrEP website tailored to youth that would increase engagement. Notwithstanding the limitations of our study, our campaign increased awareness of PrEP among some users. Because PrEP awareness and uptake is considerably low among key populations (e.g., heterosexual Black women) especially in the South, any effort that helps to increase awareness is beneficial.

This pilot study is one of the first to demonstrate the importance of using tracking pixels to monitor real-time user data to optimize performance of a digital PrEP campaign. The feedback and recommendations provided by our SMP led us to make immediate adjustments that improved engagement. The rapid pace in which digital technologies advance necessitates strong partnerships between public health practitioners and technical experts. By so doing, we will be able to fully realize the power of digital ads to advance HIV prevention efforts for young sexual and gender minorities in the South.

https://doi.org/10.1371/journal.pone.0290149

Acknowledgments

We would like to thank our collaborators at the Louisiana Department of Health, Beau Braddock (Research Coordinator), and our study participants.

Author Contributions

Conceptualization: Manuel A. Ocasio, M. Isabel Fernandez.
Funding acquisition: Manuel A. Ocasio, M. Isabel Fernandez.
Methodology: Manuel A. Ocasio, M. Isabel Fernandez, Steven Cortese.
Project administration: Steven Cortese.
Writing – original draft: Manuel A. Ocasio, M. Isabel Fernandez, Kathryn Kampa.
Writing – review & editing: Manuel A. Ocasio, M. Isabel Fernandez, Steven Cortese, Kathryn Kampa.

References

1. CDC. Diagnosis of HIV infection in the United States and dependent areas, 2021. HIV Surveillance Report 2022;34. https://www.cdc.gov/hiv/library/reports/hiv-surveillance.html.
2. CDC. Health Considerations for LGBTQ Youth: Terminology 2022. https://www.cdc.gov/healthyyouth/terminology/sexual-and-gender-identity-terms.htm.
3. CDC. Diagnoses of HIV Infection in the United States and Dependent Areas 2020: Special Focus Profiles. 2020 [updated May 24, 2022.; cited 2023 April 10]. https://www.cdc.gov/hiv/library/ reports/hiv-surveillance/vol-33/content/special-focus-profiles.html.
4. Louisiana Department of Health. 2019 STD/HIV Surveillance Report. New Orleans, LA: Office of Public Health, 2019.
5. HHS. What is Ending the HIV Epidemic in the U.S.? [updated June 2, 2021; cited 2023 April 13]. https://www.hiv.gov/federal-response/ending-the-hiv-epidemic/external.
6. Riddell J, Amico KR, Mayer KH. HIV Preexposure Prophylaxis: A Review. JAMA. 2018;319(12):1261–8. https://doi.org/10.1001/jama.2018.1917 PMID:29584848.
7. Hosek SG, Landovitz RJ, Kapogiannis B, Siberry GK, Rudy B, Rutledge B, et al. Safety and Feasibility of Antiretroviral Preexposure Prophylaxis for Adolescent Men Who Have Sex With Men Aged 15 to 17 Years in the United States. JAMA Pediatr. 2017;171(11):1063–71. https://doi.org/10.1001/jamapediatrics.2017.2007 PMID:28873128.
8. Hosek SG, Rudy B, Landovitz R, Kapogiannis B, Siberry G, Rutledge B, et al. An HIV Preexposure Prophylaxis Demonstration Project and Safety Study for Young MSM. J Acquir Immune Defic Syndr. 2017;74(1):21–9. https://doi.org/10.1097/QAI.0000000000001179 PMID:27632233.
9. CDC. PrEP Effectiveness 2022 [updated June 6, 2022; cited 2023 April 10]. https://www. cdc.gov/hiv/basics/prep/prep-effectiveness.html.
10. Reif S, Cooper H, Warren M, Wilson E. HIV in the US Deep South: Trends from 2008–2019. Durham, NC: Duke University, 2021.
11. Sullivan P, Woodyatt C, Koski C, Pembleton E, McGuinness P, Taussig J, et al. A data visualization and dissemination resource to support HIV prevention and care at the local level: analysis and uses of the AIDSVu Public Data Resource. J Med Internet Res. 2020;22(10). https://doi.org/10.2196/23173 PMID:33095177

12. Mayer KH, Agwu A, Malebranche D. Barriers to the Wider Use of Pre-exposure Prophylaxis in the United States: A Narrative Review. Adv Ther. 2020;37(5):1778–811. Epub 20200330. https://doi.org/10.1007/s12325-020-01295-0 PMID:32232664.

13. Golub SA. PrEP Stigma: Implicit and Explicit Drivers of Disparity. Curr HIV/AIDS Rep. 2018;15(2):190–7. https://doi.org/10.1007/s11904-018-0385-0 PMID: 29460223.

14. Watson RJ, Chang CJ, Feinstein BA, Moody RL, Caba A, Eaton LA. PrEP Stigma and logistical barriers remain significant challenges in curtailing HIV transmission among Black and Hispanic/Latinx cisgender sexual minority men and transgender women in the US. AIDS Care. 2022;34(11):1465–72. Epub 20220717. https://doi.org/10.1080/09540121.2022.2098908 PMID: 35848490.

15. Louisiana Health Hub. TelePrEP Program [cited 2023 April 12]. https://www.louisiana healthhub .org/teleprep/.

16. Wong KYK, Stafylis C, Klausner JD. Telemedicine: a solution to disparities in human immunodeficiency virus prevention and pre-exposure prophylaxis uptake, and a framework to scalability and equity. Mhealth. 2020;6:21. Epub 20200405. https://doi.org/10.21037/mhealth.2019.12.06 PMID:32270013.

17. Perrin A, Anderson M. Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018. Pew Research Center. 2019. https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/.

18. Swendeman D, Arnold EM, Harris D, Fournier J, Comulada WS, Reback C, et al. Text-Messaging, Online Peer Support Group, and Coaching Strategies to Optimize the HIV Prevention Continuum for Youth: Protocol for a Randomized Controlled Trial. JMIR Res Protoc. 2019;8(8):e11165. Epub 20190809. https://doi.org/10.2196/11165 PMID:31400109.

19. Ingram M. How Google and Facebook Have Taken Over the Digital Ad Industry. Fortune. 2017 [cited 2023 April 10]. https://fortune.com/2017/01/04/google-facebook-ad-industry/.

20. Goldman R. Hard Questions: What Information Do Facebook Advertisers Know About Me? Facebook 2018 [cited 2023 April 13]. https://about.fb.com/news/2018/04/data-and-advertising/.

21. Buckingham L, Becher J, Voytek CD, Fiore D, Dunbar D, Davis-Vogel A, et al. Going social: Success in online recruitment of men who have sex with men for prevention HIV vaccine research. Vaccine. 2017;35(27):3498–505. Epub 20170516. https://doi.org/10.1016/j.vaccine.2017.05.002 PMID:28526330.

22. Fontenot HB, Abuelezam NN, Rosenberger JG, Novak D, Mayer KH, Zimet G. The Impact of Advertisement Messaging on Enrollment of Young Men Who Have Sex With Men for Web-Based Research: Observational Study. J Med Internet Res. 2020;22(1):e16027. Epub 13.1.2020. https://doi.org/10.21037/mhealth-20-64 PMID:31929103.

23. Zlotorzynska M, Bauermeister JA, Golinkoff JM, Lin W, Sanchez TH, Hightow-Weidman L. Online recruitment of youth for mHealth studies. Mhealth. 2021;7:27. Epub 20210420. https://doi.org/10.21037/mhealth-20-64 PMID:33898596.

24. Dehlin JM, Stillwagon R, Pickett J, Keene L, Schneider JA. #PrEP4Love: An Evaluation of a Sex-Positive HIV Prevention Campaign. JMIR Public Health Surveill. 2019;5(2). https://doi.org/10.2196/12822 PMID:31210141

25. Braddock WRT, Ocasio MA, Comulada WS, Mandani J, Fernandez MI. Increasing Participation in a TelePrEP Program for Sexual and Gender Minority Adolescents and Young Adults in Louisiana:

Protocol for an SMS Text Messaging-Based Chatbot. JMIR Res Protoc. 2023;12:e42983. Epub 20230531. https://doi.org/10.2196/42983 PMID:37256669.

26. 2020 Census [Internet]. U.S. Department of Commerce. n.d. https://www.census.gov/programs-surveys/decennial-census/decade/2020/2020-census-main.html.

27. Auxier B, Anderson M. Social Media Use in 2021. Pew Research Center, 2021. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/.

28. Stern MJ, Fordyce E, Hansen C, Heim Viox M, Michaels S, Schlissel A, et al. Social Media Recruitment for a Web Survey of Sexual and Gender Minority Youth: An Evaluation of Methods Used and Resulting Sample Diversity. LGBT Health. 2020;7(8):448–56. Epub 20201104. https://doi.org/10.1089/lgbt.2019.0311 PMID:33147121.

29. Prescott TL, Phillips Ii G, DuBois LZ, Bull SS, Mustanski B, Ybarra ML. Reaching Adolescent Gay, Bisexual, and Queer Men Online: Development and Refinement of a National Recruitment Strategy. J Med Internet Res. 2016;18(8):e200. Epub 20160804. https://doi.org/10.2196/jmir.5602 PMID:27492781.

30. Bergin AD, Vallejos EP, Davies EB, Daley D, Ford T, Harold G, et al. Preventive digital mental health interventions for children and young people: a review of the design and reporting of research. Digit Med. 2020;3(1):133. https://doi.org/10.1038/s41746-020-00339-7 PMID:33083568

31. Koh A, Swanepoel W, Ling A, Ho BL, Tan SY, Lim J. Digital health promotion: promise and peril. Health Promot Int. 2021;36(Supplement_1):i70–i80. https://doi.org/10.1093/heapro/daab134 PMID:34897444.

32. Biello KB, Hill-Rorie J, Valente PK, Futterman D, Sullivan PS, Hightow-Weidman L, et al. Development and Evaluation of a Mobile App Designed to Increase HIV Testing and Pre-exposure Prophylaxis Use Among Young Men Who Have Sex With Men in the United States: Open Pilot Trial. J Med Internet Res. 2021;23(3):e25107. Epub 24.3.2021. https://doi.org/10. 2196/25107 PMID:33759792.

33. Lewis CF, Williams SZ, Tofighi B, Lekas HM, Joseph A, Rivera A, et al. Internet use and uptake of a web-based prevention and risk reduction intervention for persons who use drugs in New York City—WebHealth4Us study (2013–2016). Subst Abus. 2022;43(1):1119–26. https://doi.org/10.1080/08897077.2022.2028701 PMID:35499405.

34. Apple. Privacy Control 2023 [cited 2023 April 11]. https://www.apple.com/privacy/control/.

35. dashthis. View-Through Rate [cited 2023 April 10]. https://dashthis.com/kpi-examples/view-through-rate/.

36. Vibe. How do you calculate video completion rate? 2023 [cited 2023]. https://www.vibe. co/blog/how-do-you-calculate-video-completion-rate.

37. Jones JM. Social Media Users More Inclined to Browse Than Post Content. Social & Policy Issues [Internet]. 2023. https://news.gallup.com/poll/467792/social-media-users-inclined-browse-post-content.aspx.

38. Alexander L. SocialPilot. 2022. https://www.socialpilot.co/blog/facebook-ads-vs-instagram-ads#:~:text=That%20is%20not%20to%20say,Instagram%20is%20a%20better%20option.

39. Schiavo R. Health Communication: From Theory to Practice. 2nd ed: Jossey-Bass; 2013.

40. Ocasio MA, Fernandez MI, Joseph JM, Rezai R. Engaging Sexual and Gender Minority Youth in HIV Interventions Through Gay Dating Apps: Recruitment Protocol. JMIR Res Protoc. 2021;10(6):e28864. Epub 20210622. https://doi.org/10.2196/28864 PMID:34156342.

4.9 Case Study 3

RESEARCH ARTICLE

Unpacking privacy: Valuation of personal data protection

Anya Skatova [1,2,3,4]*, Rebecca McDonald [2,5], Sinong Ma[3,6], Carsten Maple[3,4]

1 Bristol Medical School/School of Psychological Science, University of Bristol, Bristol, United Kingdom, 2 Warwick Business School, University of Warwick, Coventry, United Kingdom, 3 Warwick Manufacturing Group, University of Warwick, Coventry, United Kingdom, 4 Alan Turing Institute, London, United Kingdom, 5 School of Economics, University of Birmingham, Birmingham, United Kingdom, 6 Queen's Management School, Queen's University Belfast, Belfast, United Kingdom

* *anya.skatova@bristol.ac.uk*

Abstract

Information about individual behaviour is collected regularly by organisations. This information has value to businesses, the government and third parties. It is not clear what value this personal data has to consumers themselves. Much of the modern economy is predicated on people sharing personal data, however if individuals value their privacy, they may choose to withhold this data unless the perceived benefits of sharing outweigh the perceived value of keeping the data private. One technique to assess how much individuals value their privacy is to ask them whether they might be willing to pay for an otherwise free service if paying allowed them to avoid sharing personal data. Our research extends previous work on factors affecting individuals' decisions about whether to share personal data. We take an experimental approach and focus on whether consumers place a positive value on protecting their data by examining their willingness to share personal data in a variety of data sharing environments. Using five evaluation techniques, we systematically investigate whether members of the public value keeping their personal data private. We show that the extent to which participants value protecting their information differs by data type, suggesting there is no simple function to assign a value for individual privacy. The majority of participants displayed remarkable consistency in their rankings of the importance of different types of data through a variety of elicitation procedures, a finding consistent with the existence of stable individual privacy preferences in protecting personal data. We discuss our findings in the context of research on the value of privacy and privacy preferences.

Introduction

With the rising prominence of digital technology in the modern economy, the amount of data recorded about consumers' behaviour and retained in various virtual and physical places is enormous. This personal data can reveal private details about consumers' everyday behaviour. For example, records of individuals' daily journeys are reflected in mobile phone or car GPS data; household composition can be deduced from electricity use in the home; loyalty card data can reflect personal lifestyle choices; and bank transaction data can reveal details of personal financial circumstances. These opportunities are increasingly being exploited in new and innovative ways: personal data has economic value for government organisations and businesses [e.g., in advertising, 1]. It enables a diverse collection of digital economy services to be provided, serving a foundation for many successful business models [2; see 3 for a review of the role of personal data in a variety of business models]. In this paper, we aim to understand preferences for protecting personal data.

https://doi.org/10.1371/journal.pone.0284581 Used with CC 4.0 license

Value of Personal Data and Preference for Privacy

Economists have long been interested in how personal information is shared and, broadly speaking, how it is valued (by Acquisti et al. [3], see also Stigler [4] and Poster [5]). This work brings into focus consumers' views on the consequences of their privacy-related behaviours and measures their preferences regarding privacy and value of personal data [e.g., 6–12]. An individual deciding to share data online may be compromising their privacy, since this data can reveal private information about them. Therefore, an important factor influencing an individual's preference for protecting their personal data is how much they value their privacy, captured by their preference for privacy. Privacy preferences are a feature of some theoretical models of online privacy including Montes et al. [8]. Relatedly, the well-established idea of the "privacy calculus" [13] is that individuals assess the costs and benefits of protecting (or sharing) their information—essentially asking themselves whether it is "worth it" to reveal a certain type of information. To engage in such a calculus, individuals must have defined preferences over their privacy. However, some studies raise doubts about the existence of such preferences [e.g. 14]. We review the literature regarding the existence and nature of preferences for privacy, and the value of privacy, in what follows.

An early empirical investigation into the value of privacy was undertaken by Hann et al. [15], who conducted a Discrete Choice Experiment to investigate the trade-offs between privacy and benefit (financial and time saving) in the choice between two websites. The authors demonstrated that their sample of students were willing to trade off their privacy for sufficient economic benefit. Huberman et al. [16] explored individuals' willingness to accept compensation for revealing sensitive personal information, specifically one's bodyweight. They showed that people were willing to engage in such a trade, but that those who perceived their weight to be different from the norm demanded greater compensation for sharing this data. This again suggests that individuals are entering into a form of privacy calculus.

Willingness to Pay to Protect Different Types of Personal Data

Since these early investigations, there has been ongoing interest in attempting to value the protection of privacy online [e.g., 17]. A systematic review of the literature regarding "putting a price tag on personal information" was conducted by Wagner et al. [18]. They found that the type of information being disclosed (or protected) was an important feature in determining the amount that individuals were willing to pay (or accept). The second feature that they discovered to be important was the privacy impact of personal information disclosure. Specifically, they found that:

> "the more sensitive the data and the more identifiable people are, the higher has been the price people attach to their data as they perceive higher risks" [18, p 3766].

As we have seen, there is substantial variation in the evidence about willingness to pay for protection of personal data. This may reflect the lack of any clear understanding amongst the participants—and arguably amongst the researchers—about exactly what they mean by privacy. Many of the articles refer to the value of privacy, but actually test willingness to pay not to share (or, willingness to accept compensation for sharing) a particular type of personal data [e.g. 19 for the results of a logic test; 20 for social media details]. This means that the value attributed to privacy will be confounded by all the other reasons that someone may wish to avoid sharing that particular type of personal data. Our view is that privacy relates directly to what sharing a piece of data can reveal about the person's behaviour. This will differ on many dimensions, including context, who is requesting the data, and on the specific data type. As such, to understand how individuals value their privacy, a first step is to consider their willingness to share a variety of different types of personal data.

https://doi.org/10.1371/journal.pone.0284581

However, while the values of different data types might vary, there still could be underlying individual privacy preferences. Many of the studies to date that explore the value of privacy (through eliciting the value of personal data) focus on just one type of data [e.g., demographic information, 11]. In this case, the consistency and stability of preferences for protecting personal data cannot be investigated, and the robustness of any conclusions about how much individuals value personal data in general cannot be ascertained. Exceptions to this mainly investigate the comparison between willingness to accept and willingness to pay [e.g., 12] or between different types of willingness to pay [e.g., 10] but still with the same pieces of personal information. However, different types of data—for example energy use at home versus bank transactions—can reveal different information about an individual, and these different revelations may be sensitive to different extents. There is evidence that people care about some types of data more than others, and that people differ with respect to which types of data they care about [21, 22]. Arguably, then, the value of personal data is not a unified concept with a single value but is context-dependent and will vary by different types of data [3]. Further, consumers with the same underlying privacy preferences might make different decisions while sharing different types of data because the assessment of risks and benefits will be different in each data sharing environment.

To summarise this literature, it appears that individuals are often willing to give up privacy for benefits defined in terms of money and/or convenience, and that their willingness to engage in such trade-offs may differ between contexts where the risks and benefits of the trade-offs vary. This evidence is consistent with the idea that privacy preferences exist, that they are a component of the value of personal data, and that they may influence the decision whether to share personal data online, alongside other influences related to the decision context.

However, an omission in the current literature is evidence about the stability and consistency with which individuals can express their privacy preferences. If individuals' stated preferences regarding their personal data are not consistent, stable and well-defined, then behaviourally-relevant models, in which online behaviour is governed in part by privacy preferences, will be intractable. Our paper aims to understand preferences for protecting personal data through assessing individuals' rankings of the importance of protecting different types of personal data, including (but not limited to) analysing their willingness to pay to protect them. We next describe in more detail existing methodological work on the stability of privacy preferences, and draw on related literature in other contexts, to ascertain how best to capture the stability—or otherwise—of preferences.

Preference Elicitation Across Contexts

To understand whether people can consistently report their preferences, it is necessary to elicit these preferences in different ways. To our knowledge just one paper has compared different methods of preference elicitation in the context of online privacy. Benndorf and Normann [10] investigated the effect of two different methods in eliciting the value of personal information, comparing a Becker-DeGroot-Marshak style second price auction [23] with a Take it or Leave it approach. They found a significant difference in values between these approaches, even though both aimed to elicit the same underlying construct. Hence, they provided preliminary evidence that preferences to keep personal data private—at least in terms of their monetary valuations—are subject to elicitation method effects. To our knowledge no research has yet considered the (in)consistency with which individuals express their preference ranking over the importance of protecting different types of personal data, and whether those rankings depend on elicitation methods. However, evidence about the stability and reliability of preferences expressed through different elicitation procedures has been considered in detail in other fields, most notably in experimental economics and in the non-market valuation literature applied to health outcomes.

https://doi.org/10.1371/journal.pone.0284581

Methods that exist to categorise and compare the strength of preference over different outcomes come in many forms, from simple choice tasks through to complex trade-off tasks between the outcome and some numeraire of its value [for reviews comparing methods in the health context, see 24; 25]. One important way to categorise these methods is by their relative or absolute nature. Relative tasks allow a researcher to elicit the order—and sometimes the strength of preference—for one outcome compared to another. Examples include direct ranking tasks, where a participant must rank outcomes in order of preference, and binary choice tasks where one option is compared to another and the preferred option is chosen. Absolute tasks require participants to reveal the strength of preference for an outcome by itself without reference to any alternatives. For example, with rating scales an outcome is rated in terms of its quality, likeability or importance; while in valuation tasks an outcome is judged according to its monetary value.

According to standard economic theory, as long as individuals hold well-defined and consistent preferences over the outcomes in question, the type of task used to elicit these preferences should not influence the outcome of the preference elicitation. Put simply, if I like an apple more than an orange, then I would choose an apple over an orange in direct choice, and I would be willing to pay more for the apple than I would be willing to pay for the orange. However, a long and rich literature in experimental economics and psychology has revealed that choice and judgement tasks follow different psychological processes [e.g. 26] and prioritise different attributes of a comparison [e.g. 27]. It is therefore an open question whether individuals can consistently reveal their preference ordering with respect to the importance of protecting different types of their personal data. Arguably, the more well-defined their underlying preferences, the more stable the preference ordering will be when compared across different elicitation tasks.

The health economics literature provides further evidence to support the notion that preference elicitation procedures matter. Discrepancies between values of a health outcome have been found depending on the method used to elicit them [e.g. 24, 28–30]. There is a wide array of reasons given for these discrepancies, ranging from anchoring [e.g. 31] to reference point effects and coherent arbitrariness [32]. Taken to the extreme, some authors have argued that there may not even be such a thing as underlying, stable and coherent preferences [see for example 33–35 for various discussions along these lines]. There are striking similarities between valuing and ranking the importance of different health states, and valuing and ranking the importance of protecting different types of personal data. Both health states and personal data scenarios involve making judgements and choices that involve negative emotions like fear and distress. Both scenarios are relatively unfamiliar, in the sense that few people are regularly asked to directly evaluate different scenarios regarding their health, nor their privacy. Finally, both scenarios are difficult, arguably impossible, to incentivise, yet have important implications for regulation and policymaking. The UK Treasury's Green Book guidance for policy appraisals states that

"If robust revealed preference data is not available, surveys that use willingness to pay and willingness to accept are an established alternative method known as stated preference techniques" [2, p. 42].

It seems wise, then, to draw on the lessons from health economics when considering the value of personal data and preferences to keep the data private. We put the methods used in the health economics literature to work in the context of privacy valuation, checking for the implicit ranking of types of data that they imply, and inferring the plausibility of using these elicitation techniques to establish the value of personal data to members of the public. We focus on the comparison between rankings of types of data that are based on absolute judgment (willingness to pay) tasks, and rankings of types of data based on relative judgement methods, to understand the extent to which preference orderings for personal data are stable across these frames.

https://doi.org/10.1371/journal.pone.0284581

We emphasise that we are not advocating for a market for personal data in which service users are forced to pay for online privacy. The ethical implications of this possibility are complex and their detailed exploration is outside the scope of this research. Instead, and following the rich traditions of non-market valuation in other fields like health and the environment, we are simply using Willingness to Pay as a tool for establishing the value that individuals place on their personal data. In principle, this information can feed into regulatory processes, and in particular can provide an estimate of the benefits of privacy protection measures in cost-benefit analyses. In our setting, the values are captured to allow us to explore the stability of the expressed preferences, given that those preferences are not directly observable.

In this study, **we aimed to understand preferences for protecting personal data**. Instead of assuming preferences for privacy to be captured by an exogenously determined "privacy value" constant in the utility function in all contexts, we explored how preferences differ across different data sharing environments. We did not assume that preferences are well-defined and stable but used a multitude of elicitation techniques to explore the consistency with which different types of data are ranked and valued.

Clearly, privacy is a complex issue that encompasses, and feeds into, many other related concepts. We adopt a simple conceptual approach (illustrated in Fig 4.5) in which we assume that preferences for (online) privacy are a contributor to preferences for avoiding sharing personal data, but they not the whole story. The context of the data sharing scenario will also drive a preference for avoiding sharing personal data, but is not itself part of a preference for privacy. We further consider that Willingness to Accept (or Pay) for sharing (or not sharing) personal data online is driven in part by the preference for avoiding sharing personal data, and hence by preference for privacy, but that again other considerations may play a role, such as ability to pay. At the heart of the framework is the underlying preference for privacy, and the stability of these underlying privacy preferences are necessary (although not sufficient) for the consistency of decisions about sharing personal data, and this is what we explore in our empirical work.

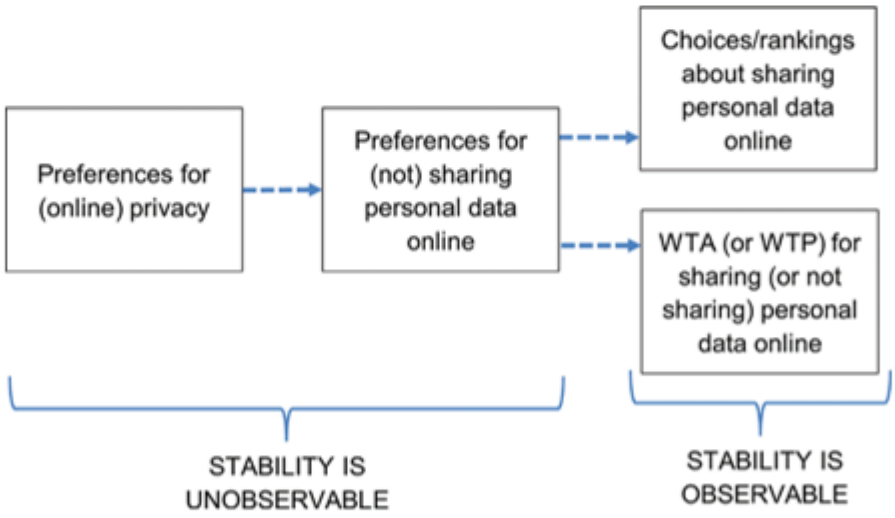


Fig 4.5 Framework for considering privacy preferences, sharing preferences, and related behaviour. Dotted arrows indicate that the components are linked, but that other influences (such as context) also contribute to determining the next component.

We took a neutral approach when defining the context in which participants were asked to report the strength of preferences for keeping their personal data private, and we did not define the specific intended use of the information they were asked to share. Our study mimics real world online environments where individuals need to make their own assessment of costs and benefits associated with sharing of a particular type of personal data, and judge for themselves whether the trade-off is worth making.

Our approach does not intend to give a detailed exploration of the reasons behind the decision whether or not to share personal data, but instead focuses on examining two main issues: how individuals value keeping personal data private in a variety of contexts and how consistently they report these preferences. We explored the reliability and stability of these preferences taking a holistic view of the prospect of sharing personal data in the context of the digital economy and compared individuals' overall willingness to share their data across personal data environments and elicited through different methods.

Our research extends previous work on factors affecting individuals' decisions about whether to share personal data. We take an experimental approach and focus on whether consumers place a positive value on protecting their data by examining their willingness to share personal data in a variety of data sharing environments, including their Willingness To Pay (WTP) to protect it. Using five evaluation techniques, we systematically investigate whether members of the public value keeping their personal data private. Specifically, we determine (1) whether stated preferences for keeping personal data private are stable within individuals; and (2) whether stated preferences for keeping personal data private systematically vary between data sharing environments (e.g., electricity consumption vs physical activity monitoring vs spending patterns).

We find that 96% of individuals are willing to pay to avoid sharing their personal data in at least one of data sharing environments, indicating that privacy is valuable. We also find that individuals are consistent in their rankings of the importance of different types of personal data when these rankings are elicited in different ways. This is evidence consistent with the notion that the techniques we employ can reveal stable underlying preferences for keeping personal data private. We find consistent differences across our sample in their stated preferences for keeping different types of personal data private: financial and medical data were typically rated as the most important to protect, while supermarket loyalty cards, electricity use and physical activity data were rated as the least important to protect. We conclude that stability in stated preferences for keeping personal data private across a variety of elicitation methods is evidence that privacy preferences are well defined. If so, the stated preferences we elicit may help to explain how individuals may decide whether to share their personal data in the context of the digital economy. Finally, we draw conclusions on the links between our findings on willingness to pay to protect personal data in different environments with the literature on the value of privacy.

Experiment

This paper investigates the stability of participants' preferences for keeping their personal data private. In an online experiment, we asked whether participants place a positive monetary value on protecting different items of their personal data and explored how their willingness to pay differed between different data sharing environments. We drew inspiration from non-market valuation economic methodologies [e.g., contingent valuation, with its origins in 36; rating scales as explored in 31; paired comparison, which draws on insights from 37], designing tasks in which participants expressed their absolute and relative judgments about the importance of protecting different items of their personal data. Through these tasks we could establish whether and how people value the privacy of their online data, whether they value different types of data differently and whether there is consistency between the rankings revealed by different elicitation methods.

https://doi.org/10.1371/journal.pone.0284581

To achieve the overall aim of understanding privacy preferences, the specific goals of our research were twofold. First, we aimed to ascertain the relative importance of each data type, as well as preliminary evidence about the levels of willingness to pay (WTP) for keeping each data type private. Second, we aimed to explore the consistency of the rankings of the data types across elicitation methods. To address these questions, we elicited the relative importance of protecting different personal data types using five different elicitation mechanisms, and gathered demographic information including age, gender, education and income to investigate whether WTP is affected by any of those factors. Two elicitation methods focused on respondents' WTP for keeping their personal data private, one further method focused on the perceived absolute importance of personal data types and two focused on respondents' ranking of the importance of protecting different types of personal data—one pairwise ranking and one direct overall ranking. This is, to our knowledge, the first study that directly addresses the effect and consistency of different elicitation mechanisms in the context of data protection.

Participants & procedure: A total of 265 participants took part in the online study run on the Qualtrics platform. The study took up to 30 minutes and participants received a fixed payment of £3 in return. The study was approved by Humanities and Social Sciences Research Ethics Sub-Committee, University of Warwick. The dataset is available upon request from the first author. Full experimental design and exact questions can be found in the S1 File.

The survey was distributed through Prolific Academic (Prolific.ac). Three participants failed to answer all questions, and 262 responses were used in all of the following analyses. Of the respondents, 69.85% were female. Participants reported an average age of 37.28 (SD = 11.42), ranging from 18 to 72. Regarding education level, 53.44% reported no undergraduate degree or equivalent, 35.11% reported an undergraduate degree or equivalent, and 11.45% reported a degree higher than undergraduate. Our sample was more educated and contained a marginally higher proportion of female participants than the general population, since the UK population is 50.89% female [38], and only 27.22% of the population have an undergraduate degree or higher qualification [39]. Income was measured on an interval scale with ten intervals: "less than £10,000"; "£10,000 to £19,999"; "£20,000 to £29,999"; "£30,000 to £39,999"; "£40,000 to £49,999"; "£50,000 to £74,999"; "£75,000 to £99,999"; "£100,000 to £149,999"; "£150,000 to £199,999"; and "£200,000 or more." There was also an option not to respond. The midpoint of each income interval was used as a proxy for the participant's income. The median of the household's annual income before tax for our sample was £25,000, equivalent to around $33,700 (SD = £23,600 or $32,000), which is lower than that of UK population in general: the Office for National Statistics [40] reported median UK household pre-tax income of £33,705 for 2017. Income levels for people who preferred not to reveal this information (12 participants) were treated as missing values.

Experimental design: The main preference elicitation section of the experimental design consisted of five conditions: two willingness to pay conditions, as well as three additional conditions to elicit individual preferences to protect different types of data. Screen shots of the five conditions are provided in Fig 4.6 and the full experimental protocol is available in S1 File. Every participant completed all five conditions, and the conditions were presented in a random order. The responses in the WTP conditions revealed whether the participant was willing to pay anything at all to protect the data types, and if so, how much they were willing to pay. The responses in the three non-WTP conditions revealed participants' ranking of data types in terms of their perceived importance of protecting the data. This allowed us to test the consistency of subjective rankings compared to the rankings implied by the WTP conditions.

https://doi.org/10.1371/journal.pone.0284581

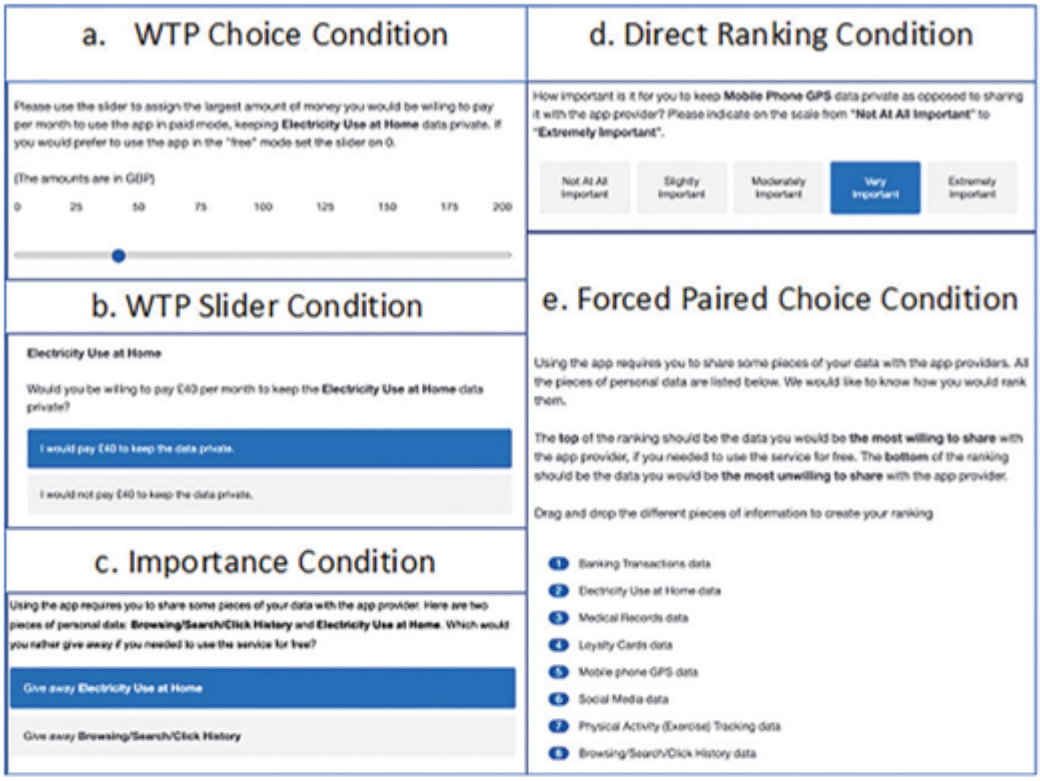


Fig 4.6 Each sub-section displays a screenshot of one of the five experimental conditions. The order of conditions was randomised betweens participants.

All conditions had a common scenario. Participants were required to imagine that they had just bought a new smartphone. In addition to having all the usual functions, it also had a special application which intended to assist them in their everyday life by tracking different types of their personal data. It was emphasised that the use of this application was very desirable for a participant. The scenario specified that when they used the personal data application, they were given two choices; they could run the app in "free" mode, or they could run it in "paid" mode. In "free" mode, they did not have to pay anything but the company providing the app collected information about them when they used the app and used that information for various marketing and other purposes. In "paid" mode, they would pay a specified amount of money for the service and their data would remain private. In "paid" mode, the data would remain anonymous in a database and would only be used to maintain the functionality of the app. We did not specify the purposes for which the data might be used in "free" mode, allowing individuals to make up their own minds about the likely uses and value of their personal data for companies.

The personal data, defined here as "*D*", could be one of the following: Banking Transactions, Browsing/Search/Click History, Electricity Use at Home, Loyalty Cards, Medical Records, Mobile Phone GPS, Physical Activity (exercise) Tracking and Social Media. These eight types of data were selected from a larger set of data types based on previous research (Skatova et al. 2013) as well as because they are commonly used in real life and are familiar to the general public.

https://doi.org/10.1371/journal.pone.0284581

WTP choice condition Participants were first asked whether they were willing to pay anything, even a small amount, to keep the data D private, or alternatively whether they would prefer to use the app for free and share their data D. If they preferred to pay, we asked them whether they would be willing to pay Great British Pounds (£) 40 per month to keep the data private or else use the app for free and share the data. Based on their response, different amounts of money were presented, and respondents had to accept or reject paying that amount to keep the data private.

The amounts of money were iterated between the bounds £0.01 and £200 until the person was indifferent. The iteration procedure was designed to reveal bounds on the individual's indifference point, which was then coded as the midpoint of these bounds. For instance, if an individual sequentially confirmed they would be willing to pay £40, £125 and £175 per month to keep data D private but they would not be willing to pay £200, then their highest acceptable amount is £175 per month while their lowest unacceptable amount is £200 per month. This individual's WTP would be recorded as £187.50 per month. If a participant indicated that they would be willing to pay £200, their indifference point could not be determined by the iteration. There were ten cases where participants were willing to pay more than £200 to protect some of their data types. In these cases, we asked them to state the maximum amount they would be willing to pay to keep this specific type of data private, which ranged from £210 to £300. Participants completed this procedure eight times, once per data type. The starting value of £40 and the maximum value of £200 was chosen based on pilot work with an open-ended WTP task.

WTP Slider condition Participants used a slider to directly assign the largest amount of money they would be prepared to pay to keep personal data D private. The slider ranged from £0 to £200 per month. The slider was set at a default of £40 per month to start with. This amount, and the £200 maximum, were chosen to match the £40 starting amount in the WTP Choice condition. Participants answered eight WTP slider tasks, one per data type.

Importance condition Participants used a five-point scale, ranging from "not important at all" to "very important," to rate the importance of keeping each piece of personal data D private as opposed to sharing it. Participants answered eight of these tasks, one per data type.

Direct Ranking condition Participants ranked the 8 types of personal data in order of the importance of keeping them private. They had to drag and drop types of data in a vertical ranking, where the top of the ranking was the data they were most willing to share, if they needed to use the app for free. At the bottom of the ranking was the data they were the most unwilling to share. Participants completed one ranking which included all eight data types.

Forced Paired Choice condition Participants made a series of forced paired choices.

Each choice was between two pieces of personal data, data D_1 and data D_2. Each time, they needed to decide which data type they would prefer to give away if they needed to use the app for free. Participants completed 28 of these choices, one for every unique binary pairing of the data types.

Analytic approach: We first tested whether participants were consistent in their privacy preferences by calculating how consistent the rankings of data types were between the five conditions within each participant. We then used a probit and an OLS regression to analyse whether the demographic information explained variation in whether and how much individuals were WTP to keep their data private, whether WTP depended on the types of data, and whether it depended on within-person consistency scores.

https://doi.org/10.1371/journal.pone.0284581

Individual consistency in ranking different data types. To examine consistency in individual preferences for sharing different types of data, we converted people's responses in each condition into rankings. The *Direct Ranking* condition data was used without conversion. For the *WTP Choice* and *WTP Slider* conditions we assumed that a higher WTP for a data type indicated a lower willingness to share. For the *Importance* condition we assumed higher-rated importance indicated a lower willingness to share. For the *Forced Paired Choice* condition we counted the number of times each data type was chosen as the one to share, and assumed that a lower count implied lower willingness to share the data. This way, if banking data was never given away it would have the lowest willingness to share and thus the highest importance rank, and if loyalty cards data was always given away, it would have the highest willingness to share and the lowest importance rank. For any ties in rankings, standard competition ranking was employed to determine the Willingness to Share (WTS) rank. This was converted to an importance score where the number increases with importance of the data type. S1 Appendix gives an example.

To measure consistency of each individual's ranking across different conditions we adopted Krippendorff's coefficient [41].

The alpha coefficient is defined as [41, page 222]:

$$a = 1 - D_0/D_e$$

where D_0 is a measure of the observed disagreement and D_e is a measure of the disagreement that can be expected by chance. In our context, the observed disagreement refers to the individual participants inconsistency in rankings for the same data type across different conditions. The coefficient ranges from −1 to 1: when individuals' sensitivity rankings are completely consistent across elicitation methods and inconsistent choices are completely absent, $D_0 = 0$ and $a = 1$. When consistent choice and inconsistent choices are purely driven by chance, $D_0 = D_e$ and $a = 0$. Alpha less than zero implies systematic differences in ratings that are not due to random chance, for example if two methods elicited opposite ratings for the data types.

Regressions. To estimate whether demographic and control variables influenced the likelihood of being willing to pay something greater than zero to protect personal data, we used mixed-model probit regressions with a random effect to account for the fact that each participant provided eight answers in each of the two WTP conditions. The Banking Transactions data type was used as a reference category. The following model was specified:

$$prob(y_i = 1) = F(\beta_0 + \beta_1 T_i + \beta_2 K_i + D_i\gamma + \varepsilon_i)$$

where Y_i is a binary variable taking the value 1 if the individual chose to pay money to protect the data, and 0 otherwise; D_i' is a vector of demographic variables (age, gender, education and income); K_i is the individual Krippendorff's alpha coefficient [41] and T_i is the type of data. The function F is the cumulative distribution function of the standard normal distribution.

To estimate whether demographic and control variables influenced *the level of WTP to keep data private* we used mixed-model ordinary least squares (OLS) regressions, again with a random effect of a participant [e.g. 42,43]:

$$y_i = (\beta_0 + \beta_1 T_i + \beta_2 K_i + D_i'\gamma + p_i'\delta + \varepsilon_i)$$

where Y_i is an individual's WTP for keep their data private and is treated as continuous; D_i', T_i, K_i, are defined as in the probit model.

https://doi.org/10.1371/journal.pone.0284581

Results

Are People Consistent in their Judgements about Different Types of Personal Data Between Conditions?

We calculated average rankings for each data type in each condition, pooling across participants. The aggregate ranking of each data type between different conditions was similar, with the exception of Mobile Phone GPS: see Fig 4.7 for aggregate ranking scores and corresponding 95% Confidence Intervals, and Fig 4.8A–4.8C for density distributions of ranks for each data type and each condition.

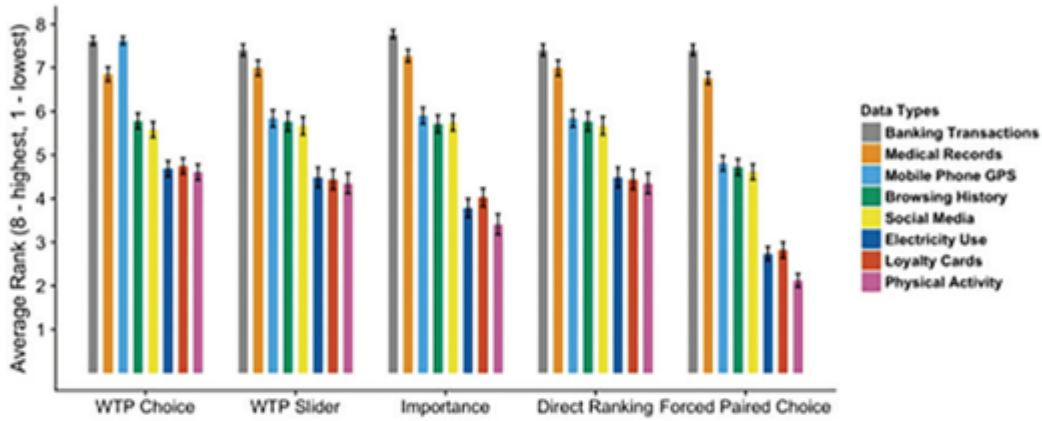


Fig 4.7 Average ranking and corresponding 95% Confidence Intervals for all conditions.

Fig 4.7 demonstrates three clear "tiers" of data. Banking Transactions and Medical Records, which we determine to be "Tier 1" data, are ranked consistently as the most valuable between all conditions, and ranked on average as 7.52 and 6.98, out of 8, respectively. A second tier of data types includes Browsing History, Mobile Phone GPS and Social Media Data, which are consistently ranked just under 6, with the exception of a lower ranking implied by the Forced Paired Choice condition for all three data types, and higher ranking for Mobile Phone GPS data in the WTP Choice condition. Finally, the lowest importance category, "Tier 3," of data includes Electricity Use, Loyalty Cards and Physical Activity data, rankings for which vary between 2 and 5 across all conditions. On average, the ranking of different data types appears consistent between conditions, with the exception of Mobile Phone GPS which is ranked in the first tier in WTP Choice but not in other conditions.

The Fig 4.8 show that Tier 1 data types are the most consistently ranked between different conditions, with density distributions that have similar shapes within the data type. Tier 2 data types have more variation and Tier 3 types have the least stability in ratings between conditions which is observable from different shapes of density distributions between conditions for the same data type. In other words, for Tier 1, and to a certain extent for Tier 2, it did not matter which elicitation procedure was used to elicit the ranks of the personal data types; the implied rankings were stable and consistent between conditions. However, for Tier 3, the shape of distribution seems to be mostly driven by the condition, not by data type: the distributions are more similar reading up the columns than along the rows for these data types. This can be interpreted to mean that the preferences with respect to Loyalty Cards, Electricity Use and Physical Activity data were less well defined, and hence the responses were driven more by the elicitation procedure than by true underlying privacy preferences.

https://doi.org/10.1371/journal.pone.0284581

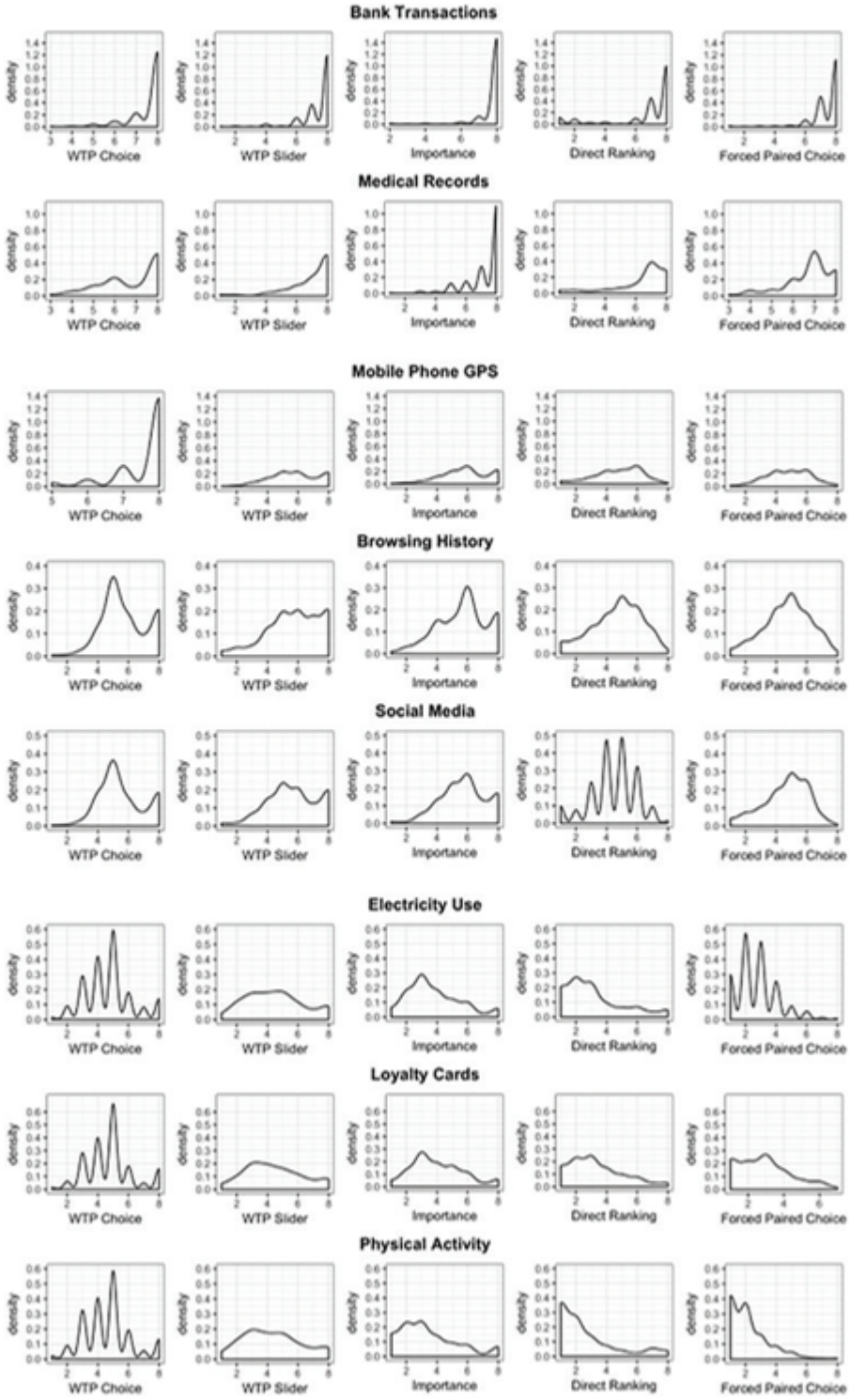


Fig 4.8 a. Density distribution of ranks for Tier 1 data: Banking and Medical Records Data. b. Density distribution of ranks for Tier 2 data: Browsing History, Mobile Phone GPS and Social Media Data. c. Density distribution of ranks for Tier 3 data: Electricity Use, Loyalty Cards and Physical Activity Data.

https://doi.org/10.1371/journal.pone.0284581

Individual level consistency: For each participant we calculated Krippendorff's alpha coefficient to reflect how consistent they were in ranking different data types. The average Krippendorff's alpha coefficient was 0.51 (median 0.60, standard deviation of 0.29), ranging from −0.17 to 0.92. For the for the degree of consistency implied by these coefficients, see rankings provided by a randomly selected participant from the bottom 5th percentile of the consistency distribution, and by a randomly selected participant in the top 5th percentile in the S2 Appendix. To describe the whole sample in terms of consistency, we adopted the benchmark scales suggested by Landis and Koch [44, see 45] which allows participants to be classified in terms of their consistency scores. The share of participants who display poor, slight, fair, moderate, substantial and almost perfect consistency are shown in Table 4.10. The table demonstrates that only 20.6% of participants are below the threshold for "fair" consistency with the majority (70.61%) being moderately, substantially or almost perfectly consistent between conditions.

Table 4.10 Break down of Krippendorff alpha consistency scores by poor, slight, fair, moderate, substantial and almost perfectly consistent.

Coefficient	Interpretation	Number of Participants	% of participants
<0	Poor	24	9.15%
[0.00–0.20]	Slight	30	11.45%
[0.21–0.40]	Fair	23	8.79%
[0.41–0.60]	Moderate	52	19.85%
[0.61–0.80]	Substantial	106	40.46%
[0.81–1.00]	Almost Perfect	27	10.30%

https://doi.org/10.1371/journal.pone.0284581.t001

To pay or not to pay? 96% of participants stated for at least one data type that they would pay to protect it. There were differences in the likelihood with which individuals would be willing to pay for different data types. Fig 9 displays the results of a Probit regression predicting whether participant were willing to pay to protect their personal data, as a function of data types, consistency and demographic variables. We used a Mixed-model Probit (predicting willingness to pay) based on the Choice condition and the Slider condition, Full regression output is in S3 Appendix.

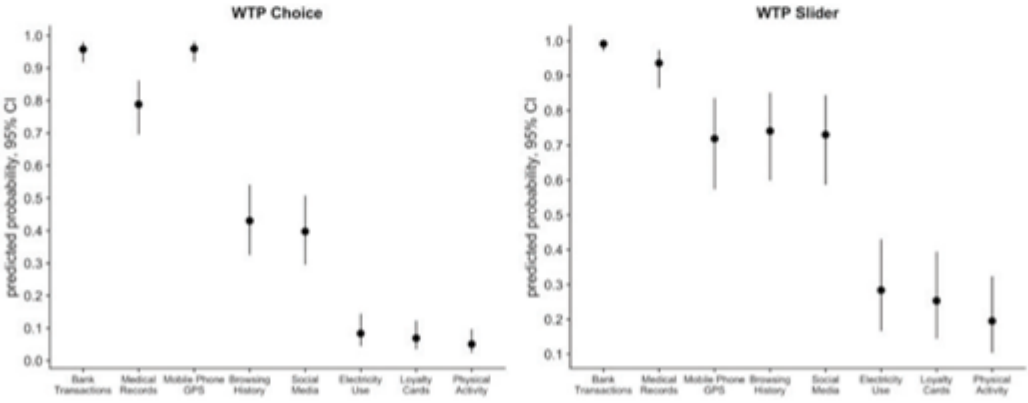


Fig 4.9 Predicted probability to pay for different data types, 95% Confidence Intervals.

The figure demonstrates predicted effects for each data type, conditioned on all other variables. The data type that was considered the most important to protect was Banking Transactions, with an estimated likelihood of being willing to pay to protect at 95.8% in WTP Choice and 99.2% in WTP Slider conditions, respectively. For Medical Records, the estimated likelihood of being willing to pay to protect was 78.9% in WTP Choice and 93.6% in the WTP Slider conditions, respectively. The second tier of data are Browsing History (43% in WTP Choice and 74.1% in WTP Slider) and Social Media (39.8% in WTP Choice and 73.1% in WTP Slider). Mobile Phone GPS was closer to the first tier in the WTP Choice condition with 96% predicted probability to pay, while in WTP Slider condition it was less protected: 71.9%. Finally, the third tier included Loyalty Cards (6.9%, WTP Choice; 25.3%, WTP Slider), Electricity Use (8.4%, WTP Choice; 28.4%, WTP Slider) and Physical Activity (5.1%, WTP Choice; 19.5% WTP Slider).

Individual consistency between ranking orders predicted higher likelihood of being willing to pay to protect personal data: those participants who demonstrated higher consistency in their responses were more likely to decide to pay something in both WTP conditions—that is, a one standard deviation increase in people's consistency measurement is estimated to cause a 9% and a 4% increase in the likelihood to pay to protect their personal data in the WTP Choice and WTP Slider conditions, respectively. This might reflect more thorough consideration of consequences of sharing personal data by those who showed higher consistency.

Out of the demographic variables, only age was negatively related to the likelihood to pay to keep the data private, and this was the case only in the WTP Slider condition: younger people were more willing to pay to keep their data private, with a one standard deviation (~11.42 years) decrease in age leading to a significant increase of 6% in the likelihood of being willing to pay to keep data private. Further, more educated participants, specifically those who stated that they held a postgraduate degree, were 10% more likely to choose to pay in WTP Choice condition. Income level had no significant effect on this decision.

See Table A2 in S3 Appendix for the regression output from the Mixed-model Probit (predicting whether participant were willing to pay for the relevant conditions.

How much to pay? We have considered the results of the Probit analysis which investigated the factors influencing whether or not participants were willing to pay a non-zero amount to protect their personal data. Next, we turn to the OLS regression which investigates the level of WTP, conditional on being willing to pay an amount greater than zero. Fig 10 displays the predicted WTP amounts in pounds and the associated 95% Confidence Intervals for each data type based on the results of the OLS regression. These results differed from, but did not seriously contradict, the results of the Probit analysis. We discuss the OLS results and differences from the Probit below.

The results related to the differences between data types were consistent with those from the Probit analysis, suggesting—reassuringly- that the data types that people are more likely to pay to protect are also the data types that they are willing the pay largest amounts to protect. The estimated WTP to protect Bank Transaction data was £28.10 per month in the Choice condition and £22.80 in the Slider condition. The estimated WTP for protecting Medical Records was £25.10 for the Choice condition and £22.50 for the Slider condition. The estimated WTP for Slider and Choice conditions for the Tier 2 data types were as follows: Mobile Phone GPS £28.50 and £11.80; Browsing History £12.12 and £12.60; and Social Media £10.10 and £11.40, respectively. Finally, our participants were willing to pay £6.61 and £7.27 for Electricity Use, £6.58 and £7.74 for Loyalty Cards and £5.95 and £7.17 for Physical Activity data in the WTP Choice the WTP Slider condition, respectively.

https://doi.org/10.1371/journal.pone.0284581

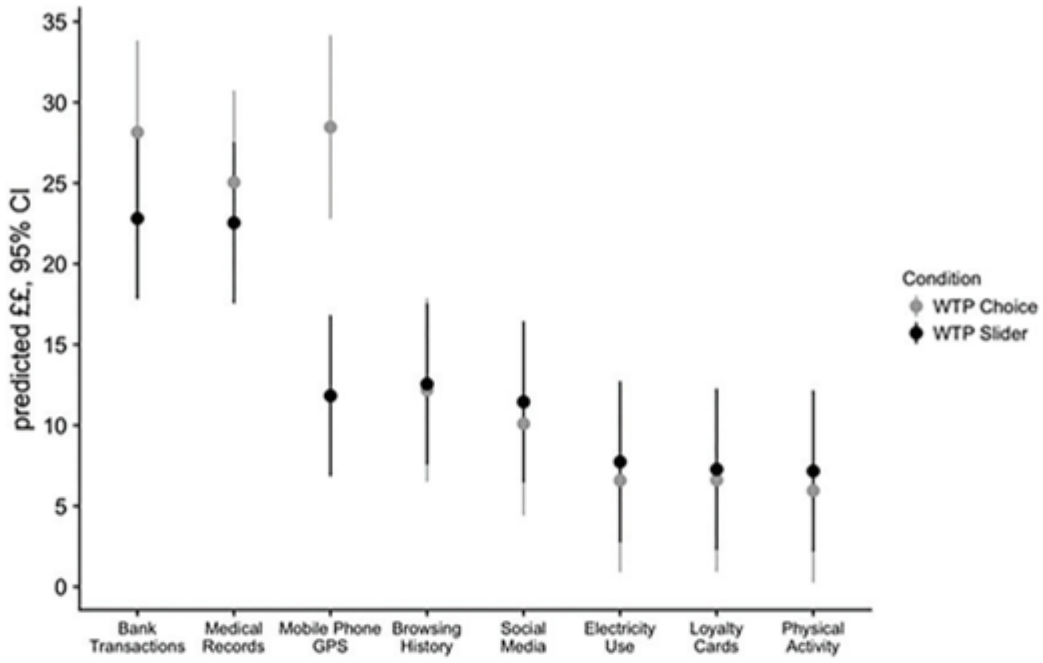


Fig 4.10 Predicted amount individuals are willing to pay for different data types, out of those who were prepared to pay something, 95% Confidence Intervals.

Individual consistency in ranking order within condition was not associated with the amount of money participants were willing to pay in either condition. Similar to the decision about whether to pay anything at all, the amount that participants were willing to pay was associated negatively with age: one standard deviation decrease in years (~11.42 years) leads to an average payment to protect personal data that is £2.27 higher, but only in the WTP Slider condition. Female participants were willing to pay smaller amounts compared to men in both WTP conditions: £8.06 less in Choice and £5.13 less in Slider, on average. Finally, having a graduate degree was associated with being willing to pay more to protect data in both conditions: £18.71 more in the Choice condition and £10.06 more in the Slider condition. As with the Probit, income did not have a significant effect on WTP. See Table A2 in S3 Appendix for the regression output OLS model predicting the amount people were WTP,

Discussion

This paper makes three main contributions. First, we showed that individuals state they are willing to pay non-zero amounts to protect their personal data. Second, individuals demonstrate consistent rankings of the importance of protecting different types of their personal data, with values and ranks being remarkably similar across different elicitation methods. Third, different types of personal data are valued differently, suggesting there is no unified "price tag for privacy." Our findings reveal consistency in people's decisions about whether to share their personal data through a variety of elicitation procedures, which is evidence for consistent with the notion of stable preferences for privacy.

We now discuss how our findings relate to previous research exploring the value of personal data, explain how our methodology allowed us to unpack privacy and its value, and discuss how our results shed light on the possibilities for emerging business models and policy approaches in the domain of personal data protection. Different types of personal data might have different value for different stakeholders, and personal

https://doi.org/10.1371/journal.pone.0284581

data may even have different value for the same consumer in various scenarios (e.g., a scenario in which a company provides services in exchange for consumer data vs a scenario in which consumers sell their data to company). Our study is the first to unpack the relative importance of protecting different types of personal data, ranging from very subjectively important, such as medical records and financial information, to—at the present time—subjectively unimportant, such as loyalty cards and electricity data. Our study deliberately did not control for specific risks and benefits of sharing as we aimed to mimic the real-world choices that people often face, where risks and benefits of agreeing to share one's data are typically unclear, although future research could usefully contribute to the literature by exploring these types of effect.

Privacy Preferences

Our work extends the growing literature on the monetary valuation of privacy [16, 46]. It also contributes to the body of knowledge concerning the factors that influence individuals' decisions about whether or not to share elements of their personal data online. To the best of our knowledge, this is the first report of a study that directly compares the value of protecting different types of personal data using five different elicitation techniques. We explored two main questions: whether individuals would state that they were willing to pay to protect different types of personal information, and whether their responses would imply rankings that were stable across different evaluation techniques, since economic theory suggests that if stable and well-defined privacy preferences exist, individuals' stated WTP would be consistent across different evaluation techniques.

In line with some previous research [18] and in contrast to other studies [46], our findings show that individuals state positive willingness to pay to ensure their personal data remains private, at least for some types of data. We observed a surprising degree of consistency in participants' rankings of different data types elicited through different techniques, which indicates that stated preference evaluation techniques may be appropriate tools to reveal individuals' underlying preferences for privacy. In our study, all factors that could affect people's decisions about what value they assign to different types of personal data remained the same, with only data types and evaluation techniques varying between the questions. We suggest that while the rankings implied by the responses in our study display remarkable consistency, there are still features of the decision context that could influence the value of privacy online. These features may include the context of data sharing, the type of data that is shared, what the purpose of sharing the data is, and who the data is shared with. We suggest that these features will interact with underlying privacy preferences and contribute to producing the observed values of personal data. By not specifying or deliberately manipulating these features of the data sharing environment in our scenarios, we examined the effects of data type and elicitation procedure in isolation.

Our methodological approach and results demonstrate the benefits of clearly specifying privacy decision-making scenarios as choices about protecting specific types of data. The clear differences in value between different data types suggests that the value of privacy in the digital world cannot be defined as a unified concept: it is not possible to assign a single price tag to individual privacy. Instead, our findings suggest that in order to understand the value of privacy, it is necessary to study how individual privacy preferences play out in different data sharing environments. For example, in our study there was a clear contrast between very high rates of willingness to pay to protect banking transactions data and very low rates of willingness to pay to protect electricity use or physical activity data, demonstrating that public preferences for keeping the former type of data private are stronger than for the latter types. In addition, responses about low-valued data, referred to earlier in the paper as Tier 3, were noisier and appeared to be more strongly influenced by the elicitation techniques than for the high-valued Tier 1 and 2 data types. This result suggests that individuals have less well-defined preferences about protecting the Tier 3 data types than they do for the highly valued data types.

https://doi.org/10.1371/journal.pone.0284581

Understanding how much people value different types of personal data is an important first step in understanding individual privacy preferences and how they affect decisions in the real world. Everyday, people face scenarios where they are presented with requests to share different types of personal data, albeit the context might vary. Our results and the evaluation framework that we have introduced in this paper could be used to further investigate how those valuations change with different framing [e.g., similar to 47], or when the instructions manipulate the emphasis placed on the risks versus the benefits of sharing personal data online. However, we emphasise again that we are not advocating for the creation of markets for personal data in which individuals are required to pay to protect their privacy. Whilst our results suggest that at least some people would be willing to engage in such a market, the ethical considerations of this approach to managing online personal data are complex and the potential for exacerbation of inequality, and for the emergence of privacy as a luxury good, are troubling. Instead, by demonstrating that individuals' willingness to pay to protect their personal data is positive and relatively stable, we hope to have provided valuable information about the strength of public preferences over privacy, which can feed into the development of online protections, regulations, and business models that better reflect the preferences of the public. Essentially, by placing a monetary value on the protection of personal data, we provide an estimate for the benefit side of a benefit cost analysis of privacy protection.

Limitations and Future Research

The choices in our experiments were hypothetical, following the tradition of the literature on the monetary valuation of non-market goods. We opted to take a hypothetical approach because it was not possible to identify real life choices (revealed preferences) in which different data types are traded in otherwise identical scenarios. Whilst incentivizing the sharing of personal data in experimental conditions is possible, and has been done before [e.g., 46], it was not appropriate for our research question because it is not possible to conduct an ethical experiment where individuals would be requested to pay money to avoid sharing their personal data in a variety of highly sensitive data sharing environments. Using hypothetical stated preference tasks avoids this necessity by allowing us full control of the data sharing and payment scenarios and making it possible to ask about sensitive data types.

Furthermore, it is not currently feasible to construct an incentivised study where these different types of personal data are traded. Based on our findings, which showed remarkable consistency in implied rankings of and valuations for privacy, future research could be usefully conducted that focuses on a single type of data (e.g., retail loyalty cards) involving a field study where individuals could opt to pay for privacy or else opt to benefit from a free service. Building such a scenario would allow the researcher to manipulate different aspects of context of data sharing, to examine whether and how these aspects affect individual decisions. It would unlikely, however, to conduct such experiment with multiple different data types.

Decisions about sharing personal data are always contextual [3]. In our study we deliberately did not define context and reasons for sharing personal data. Future research could manipulate whether the data is shared for commercial purposes or for the public good to investigate whether this affects people's decisions about whether to share their personal data, and whether different reasons for sharing affect people in a different way.

We might also consider willingness to accept instead of willingness to pay: the literature on the economics of privacy and security implicitly assumes that the market behaviour of individuals for both actions is identical, but related literature in psychology suggests otherwise. Some previous research [e.g.,11] elicits individuals' willingness to accept money in return for sharing their personal data, instead of willingness to pay to protect it. This approach can underpin business models in which individuals have the right to privacy but can choose to sell their personal data, for example through companies like digi.me that give individuals the chance to directly choose how much data to share for financial reward. In contrast, we took a willingness to pay approach and found that individuals' willingness to pay to protect their data was positive. Indeed, since the

introduction of the GDPR legislation in the EU, the status quo has shifted to one where the data subject has the right to control their data which might make a WTA framework more appropriate in future analyses. Nonetheless, a WTP approach is in line with most non-market valuation studies across contexts [48]. A final alternative, that might be explored in future research, is one in which consumers receive a specific, non-monetary benefit for sharing their data. Understanding the gap between willingness to pay and accept, as well as whether different scenarios affect values which are produced through different elicitation techniques, may have implications for future business models which evolve based on the use of personal data.

Implications

Finally, our results have implications for policy-making. Currently, the digital economy does not typically provide a direct monetary benefit for consumers who create personal data. The benefits are instead provided through increased convenience of service delivery, or better opportunities for accessing certain services. Largely, this relies on the assumption that the actual value of one consumer's personal data is very hard to estimate and even if it is possible, the value is so small that consumers do not consider their personal data to be valuable enough to demand monetary compensation for sharing it. However, our findings, in line with previous research [see 3] suggest that there may exist an imbalance in power between the consumer and the data collector/holder in terms of the allocation of the welfare gain, whereby the consumer places a positive value on protecting their personal data and yet the welfare gain from the use of that personal data is allocated solely to the data handler.

Our results strongly suggest that privacy is not a unified concept and people make different decisions depending on what kind of data they are sharing. The types of detriment that are experienced when data are not protected can be diverse—including privacy loss *per se*, risks of adverse consequences for the individual, and a basic disapproval of situations where a private company makes money from one's own personal data. It is important to understand the preferences of members of the public towards all of these elements when making decisions about how to create and regulate a fair environment in which individuals are able to make informed choices about protecting their privacy. Similar to health economics where resources are distributed based on public evaluations of different health outcomes [2], arguably government regulations with respect to personal data and digital privacy should take account of the value of personal data privacy to society. Our study is the first to look at a breadth of different data types in terms of how they are evaluated by members of public, and we provide reassuring evidence in favour of the stability of preferences for online privacy. Our results therefore indicate how public preferences might be included in policy making in this important context.

Supporting Information

S1 Appendix. Example of converting forced paired choices into a ranking.
(DOCX)

S2 Appendix. Two randomly selected participants from the top and bottom consistency scores distribution.
(DOCX)

S3 Appendix. Tabular format of regression output.
(DOCX)

S1 File.
(DOCX)

Author Contributions

Conceptualization: Anya Skatova, Rebecca McDonald, Carsten Maple.
Data curation: Anya Skatova, Rebecca McDonald.
Formal analysis: Anya Skatova, Rebecca McDonald, Sinong Ma.
Funding acquisition: Anya Skatova, Carsten Maple.
Methodology: Anya Skatova, Rebecca McDonald, Carsten Maple.
Writing – original draft: Anya Skatova, Rebecca McDonald, Sinong Ma, Carsten Maple.
Writing – review & editing: Anya Skatova, Rebecca McDonald, Sinong Ma, Carsten Maple.

References

1. Shen Q, Miguel Villas-Boas J. Behavior-based advertising. Management Sci, 2017; 64(5), 2047–2064.
2. HM Treasury Green Book (2018) The economic value of data: discussion paper. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/731349/20180730_HMT_Discussion_Paper_-_The_Economic_Value_of_Data.pdf
3. Acquisti A, Taylor C, Wagman L. The economics of privacy. J. Econ. Lit. 2016; 54(2), 442–92.
4. Stigler GJ. An introduction to privacy in economics and politics. J. Legal Stud. 1980; 9(4), 623–644.
5. Posner RA. Economic theory of privacy. *Regulation*. 1978; 2, 19.
6. Acquisti A, Brandimarte L, Loewenstein G. Privacy and human behavior in the age of information. Science 2015; 347(6221), 509–514. https://doi.org/10.1126/science.aaa1465 PMID:25635091
7. Casadesus-Masanell R, Hervas-Drane A. Competing with privacy. Management Sci, 2015; 61(1), 229–246.
8. Montes R, Sand-Zantman W, Valletti TM. The value of personal information in online markets with endogenous privacy. Management Sci. 2019; 65(3), 1342–1362.
9. Spiekermann S, Korunovska J. Towards a value theory for personal data. J. Inf. Technol. 2017; 32(1), 62–84.
10. Benndorf V, Normann HT. The willingness to sell personal data. *Scand. J. Econ*. 2017; 120(4), 1260–1278
11. Beresford AR, Kübler D, Preibusch S. Unwillingness to pay for privacy: A field experiment. Econ. Lett. 2012; 117(1), 25–27.
12. Acquisti A, John LK, Loewenstein G. What is privacy worth? J. Legal Stud. 2013; 42(2), 249–274.
13. Laufer RS, Wolfe M. Privacy as a concept and a social issue: A multidimensional developmental theory. J. Soc. Issues. 1977; 33(3), 22–42.
14. Tucker CE. Social networks, personalized advertising, and privacy controls. J. Mark. Res. 2014; 51(5), 546–562
15. Hann IH, Hui KL, Lee TS, Png IP. The value of online information privacy: Evidence from the USA and Singapore. Internat. Conf. Information Systems. 2002
16. Huberman BA, Adar E, Fine LR. Valuating privacy. IEEE Security & Privacy 2005; 3(5), 22–25.

17. Rose EA. An examination of the concern for information privacy in the New Zealand regulatory context. Information & Management. 2006; 43(3), 322–335.

18. Wagner A, Wessels N, Buxmann P, Krasnova H. Putting a Price Tag on Personal Information-A Literature Review. Proc. Hawaii Internat. Conf. on System Sci. 2018

19. Jentzsch N. Auctioning Privacy-Sensitive Goods: A Note on Incentive-Compatibility. Privacy Technologies and Policy: Second Annual Privacy Forum, APF 2014, Athens, Greece, May 20–21, 2014, Proceedings (Vol. 8450, p. 133). Springer.

20. Schreiner M, Hess T. Why Are Consumers Willing to Pay for Privacy? An Application of the Privacy-freemium Model to Media Companies. European Conf. on Information Systems. 2015; Paper 164.

21. Skatova A, Johal J, Houghton R, Mortier R, Bhandari N, Lodge T, et al. Perceived risks of personal data sharing. Proc. Digital Economy: Open Digital. 2013

22. Goldfarb A, Tucker CE. Shifts in privacy concerns. Am. Econ. Rev. 2012; 102(3), 349–53.

23. Becker GM, DeGroot MH, Marschak J. Measuring Utility By a Single-response Sequential Method. Behav. Sci. 1964; 226–32. https://doi.org/10.1002/bs.3830090304 PMID:5888778

24. Froberg DG, Kane RL. Methodology for measuring health-state preferences—II: Scaling methods. J. Clin. Epidemiol. 1989; 42(5), 459–471. https://doi.org/10.1016/0895-4356(89)90136-4 PMID:2732774

25. Nord E Methods for establishing quality weights for life years, National Centre for Health Program Evaluation. 1991

26. Rosen LD, Rosenkoetter P. An eye fixation analysis of choice and judgment with multiattribute stimuli. Mem. Cogn. 1976; 4(6), 747–752. https://doi.org/10.3758/BF03213243 PMID:21287006

27. Tversky A, Sattath S, Slovic P. Contingent weighting in judgment and choice. Psychol. Rev. 1988; 95(3), 371.

28. Krabbe P, Essink-Bot ML, Bonsel G. The comparability and reliability of five health-state valuation methods. Soc. Sci. Med. 1997; 45(11), 1641–1652. https://doi.org/10.1016/s0277-9536(97)00099-3 PMID:9428084

29. Hornberger JC, Redelmeier DA, Petersen J. Variability among methods to assess patients' well-being and consequent effect on a cost-effectiveness analysis. J. Clin. Epidemiol. 1992; 45(5), 505–512. https://doi.org/10.1016/0895-4356(92)90099-9 PMID:1588356

30. Salomon JA, Murray CJ. A multi-method approach to measuring health-state valuations. Health Econ. 2004; 13(3), 281–290. https://doi.org/10.1002/hec.834 PMID:14981652

31. Sutherland HJ, Dunn V, Boyd NF. Measurement of values for states of health with linear analog scales. Med. Decis. Mak. 1982; 3(4), 477–487.

32. Ariely D, Loewenstein G, Prelec D. "Coherent arbitrariness": Stable demand curves without stable preferences. Q. J. Econ. 2003; 118(1), 73–106.

33. Fischhoff B. Value elicitation: Is there anything in there? Am. Psychol. 1991; 46(8), 835.

34. Slovic P. The construction of preference. Am. Psychol. 1995; 50(5), 364.

35. Loomes G. (How) Can we value health, safety and the environment? J. Econ. Psychol. 2006; 27(6), 713–736.

36. Dupuit J. On the measurement of the utility of public works. Internat. Econ. Papers, 1944; 2, 83–110.

37. Thurstone LL. A law of comparative judgment. Psychol. Rev. 1927; 34, 273.

38. Office for National Statistics (2011a) 2011 Census: Population Estimates for the United Kingdom,http://webarchive.nationalarchives.gov.uk/20160108132257/http://www.ons.gov.uk/ons/dcp171778_292378.pdf

39. Office for National Statistics (2011b) 2011 Census: Key Statistics for England and Wales, https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/2011censuskeystatisticsforenglandandwales/2012-12-11

40. Office for National Statistics (2017) Household disposable income and inequality in the UK: Financial year ending 2016. London: Office for National Statistics. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/bulletins/householddisposableincomeandinequality/financialyearending2016.

41. Krippendorff K. Content analysis: An introduction to its methodology, 2nd edn. 2004; Thousand Oaks, CA: Sage.

42. Greene WH (1997) *Econometric analysis*, 3rd edn (Upper Saddle River, NJ: Prentice Hall).

43. Cameron AC, Trivedi PK. Microeconometrics using Stata, revised ed. 2010, College Station, TX: Stata Press.

44. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; pp.159–174. PMID:843571

45. Krippendorff K. Computing Krippendorff's alpha-reliability. 2011; Working Paper, University of Pennsylvania, Pennsylvania.

46. Grossklags J, Acquisti A. When 25 Cents is Too Much: An Experiment on Willingness-To-Sell and Willingness-To-Protect Personal Information. Workshop Econ. Information Security. 2007

47. Adjerid I., Acquisti A., & Loewenstein G. Choice architecture, framing, and cascaded privacy choices. Management Sci. 2019; 65(5), 2267–2290.

48. OMB (U.S. Office of Management and Budget). Circular A-4: Regulatory Analysis. Washington, DC, September 17, 2003.

4.10 Discussion Topics

The theme for the discussion this week is Privacy Risks & Principles. Privacy is a basic human right that is important in the realization of personal freedom, respect for the person and the protection of personal information. Due to the fact that in the current world there is a lot of interaction between people and a good deal of information is being shared across borders, there is need for well outlined laws and regulations that are going to protect people from being violated in their privacy. This forum will offer a detailed analysis of privacy in the modern digital age and will include many different perspectives that can be used to address the theme. Below are some of the points that can be used to discuss these issues in detail:

1. From Physical to Digital Privacy: What used to be about protecting physical areas now concerns protecting personal data in the digital world.
2. Social Media's Role in Redefining Privacy: How have social media networks affected people's perception and understanding of privacy?
3. Generational Differences in Privacy Concerns: Are young people less concerned about privacy in the digital age than older people? If so, why?
4. Public vs. Private in the Digital World: How do people distinguish between what is public and what is private in a world where everything seems to be shared?
5. The Trade-off Between Convenience and Privacy: Is privacy being given up for convenience when people use digital services?
6. The Consequences of Data Breaches: What are the long term impacts of large scale data breaches (e.g. Equifax, Yahoo) on individuals and companies?
7. Third-Party Data Sharing: Should companies be responsible for privacy violations that take place through third party relationships (e.g. Facebook and app developers)?
8. The Dark Web and Stolen Data: How are data breaches contributing to the growth of criminal activities on the dark web and what can be done to prevent this?
9. The Role of Human Error in Privacy Violations: How many times privacy violations and data breaches happen because of human mistakes instead of hacking attacks?
10. Government Surveillance and National Security: Is mass government surveillance (e.g. NSA, PRISM) a more effective way to enhance national security or is it an invasion of privacy of the people?
11. Corporate Surveillance and Behavioral Tracking: How do companies like Google and Facebook collect data from their users using surveillance methods and what are the moral issues surrounding this?
12. The "Chilling Effect" of Surveillance: How does the fact that one's activities are being monitored on the Internet influence the way an individual behaves or talks?
13. The Role of Surveillance Technologies (e.g., Facial Recognition): How do technologies such as facial recognition impact on privacy in public places? Should they be regulated?
14. Surveillance in the Workplace: To what extent should employers be permitted to watch their employees' digital activities at work?
15. The Role of Encryption in Safeguarding Privacy: How well are encryption tools like VPNs and encrypted messaging services in protecting people's privacy?

16. Managing Online Footprints: What measures can people take to reduce their online presence and the amount of information that companies can gather about them?
17. Legislation for Online Privacy Protection: Are laws such as GDPR and CCPA enough to protect online privacy or are we in need of more stringent legislation?
18. Awareness and Education as Key Strategies: How can we increase public awareness of privacy protection online and the rise of cyber threats?

Use the knowledge from the textbook together with your personal experience, opinion, and judgment to enhance the attractiveness of your post to the discussion.

4.11 Projects

4.11.1 Right to be Forgotten: Privacy versus Freedom of Speech in the Digital Age

Abstract: During this four-week undergraduate research project the right to privacy is studied in relation to the freedom of speech through a German court case about a convicted murderer who wanted to remove his crime details from Wikipedia after serving his prison term. The project will investigate the legal and ethical aspects of the right to be forgotten in order to see its effects on privacy and public information. The project will solve the issue by analyzing the conflict through legal research, case study analysis, and ethical reasoning to see how modern societies should strike a balance between the two competing interests.

Research Questions: What privacy rights does Germany give to rehabilitated criminals and are these rights different from those of other countries? What is the relation between the right to be forgotten and freedom of speech especially in platforms such as Wikipedia that are known for their public knowledge resources? Is it ethical to erase information about criminal cases from the public sphere once a sentence is served?

Methodology: Legal Research: This involves the examination of the privacy laws of Germany and other international law precedents related to the right to be forgotten and freedom of expression.

Case Study Analysis: The researcher will study the case of the German murderer and other cases in order to understand the legal decisions and judgments made in such cases.

Ethical Reasoning: The researcher will engage in ethical debates and reasoning exercises to analyze the moral principles of privacy and public knowledge.

Comparative Analysis: This will involve the comparison of the right to be forgotten in different jurisdictions and the effects on freedom of expression.

Expected Outcomes: A comprehensive report on the legal privacy rights for rehabilitated criminals in Germany and other countries. A well-balanced discussion of the moral dilemma between the right to privacy and the right to information of the public. Suggested changes to policies or laws that can help to achieve a balance between the right to privacy and freedom of expression in the digital environment. A set of guidelines or principles that could be useful for platforms such as Wikipedia in dealing with such conflicts.

Timeline:

- Week 1: Project Planning and Initial Research. Describe the project goals and strategies. Look at the privacy laws of Germany and the right to be forgotten in a general sense.
- Week 2: Case Study and Comparative Analysis. Focus on the Wikipedia case and collect data from other international cases that are similar to it. Begin the comparison of how various jurisdictions address privacy concerns.

- Week 3: Ethical Reasoning and Debates. Arrange for debates or group discussions on the ethical aspects of privacy rights versus freedom of expression. Combine the findings from legal research and case studies to support these discussions.
- Week 4: Reporting and Recommendations. Combine the legal, ethical and comparative analyses in a single report. Develop policy recommendations that are sensitive to privacy rights and freedom of information. Create a presentation to present the project's results and suggested answers.

Conclusion: The project will help to contribute to the global debate on privacy and freedom of speech in the digital world, and shed some light on how the right to be forgotten is handled in different jurisdictions and the ethical issues involved.

4.11.2 Digital Footprints in Employment: Privacy, Free Speech, and Legal Boundaries

Abstract: An online project spanning four weeks at the undergraduate level investigates the ethical legal implications employers face when they gather applicant data through online methods during hiring. The research investigates whether job candidates' privacy rights remain protected from employer monitoring and whether political opinion screening in hiring represents free speech violation and whether we need to establish legal controls on digital data use in employment. The project examines present-day hiring practices relative to personal privacy and freedom of speech rights across private and public sectors through legal research combined with ethical assessments and field observations.

Research Questions: Employers who seek web-based information about applicants do they violate privacy rights? A hiring decision based on political views expressed online by applicants creates free speech violations according to legal standards yet this principle differs between public and private employers. Which web-based candidate information should prospective employers not access according to law since current restrictions do not cover these data points?

Methodology: The study analyzes the existing body of law which defines employment methods and protects privacy and free speech rights. Surveys: The research includes two separate surveys for HR professionals to gauge current employment procedures and for job candidates to reveal their privacy expectations. The analysis uses ethical reasoning methods to investigate the ethical implications of employers using online candidate information during hiring. The research investigates how employment practice rules differ across jurisdictions with a particular focus on the usage of digital information in recruitment processes.

Expected Outcomes: The research presents an overview of how web information impacts hiring decisions from both ethical and legal perspectives. Survey results reveal current hiring practices together with societal views about privacy and free speech rights in employment. The report includes operational guidelines which help employers protect candidate rights and achieve organizational goals. The project develops possible legal amendments to protect job candidates' privacy together with their free speech rights while distinguishing public from private sector employment.

Timeline:

- Week 1: Project Initiation and Legal Research. The research project outlines both its specific goals and full research plan. Conduct extensive research into privacy and free speech laws and employment standards.

- Week 2: Survey Design and Distribution. The research team delivers finalized surveys to HR professionals and job candidates. The research will continue by studying the differences between private sector and public sector employment through legal research.
- Week 3: Data Collection and Ethical Analysis. Collect and compile survey responses. Ethical analysis sessions will be conducted to evaluate the moral aspects of employer-initiated web searches.
- Week 4: Data Analysis, Reporting, and Recommendations. The study combines survey data findings with legal and ethical research results for analysis. The research team should produce a report that combines legal and ethical assessment with data findings and practical solutions. Present a summary of key findings alongside proposed employment best practices during the project's presentation.

Conclusion: The research aims to clarify how employers validate job candidates while protecting their privacy rights as it contributes to developing necessary standards and legal changes during digital transformation.

4.12 Exercises

1. How has the digital age expanded the scope of privacy?
2. What role has social media played in the erosion of privacy?
3. What are some of the concerns raised by data collection in the digital age?
4. What must be balanced when considering privacy in the digital age?
5. What are key challenges for individuals when navigating the digital age in terms of privacy?
6. What are critical aspects of maintaining privacy in the evolving digital landscape?
7. In what areas has the collection of personal data been beneficial?
8. Why is the Fourth Amendment considered to be at a crossroads in the digital age?
9. What ongoing debate does the Fourth Amendment face in the context of changing technology?
10. What was the central issue in the Cambridge Analytica scandal?
11. What concerns did the Cambridge Analytica scandal raise?
12. Why did the Equifax data breach raise concerns about organizational responsibility?
13. What lesson can be learned from the Yahoo data breaches regarding user confidence?
14. What are the general risks posed by privacy violations and data breaches?
15. How have surveillance technologies impacted individual anonymity in public spaces?
16. What is the 'chilling effect' in the context of surveillance?
17. How can surveillance technologies affect personal relationships and intimacy?
18. How can surveillance technologies perpetuate discrimination?
19. Why is balancing security and privacy important in surveillance practices?
20. Why is using strong and unique passwords important for online privacy?
21. What does Two-Factor Authentication (2FA) add to account security?

22. Why are regular software updates important for online privacy?
23. How can adjusting privacy settings on social media platforms help protect your personal information?
24. What is the role of encryption in online privacy?
25. Why is enforcing privacy laws a challenge in the digital age?
26. What privacy concerns are raised by emerging technologies such as facial recognition and artificial intelligence?
27. What is the role of transparency in privacy laws?

Chapter

5

Cybercrime & Security

- Types of cybercrime and their impact on society
- Methods used to prevent and detect cybercrime
- Ethical considerations surrounding cybersecurity
- The role of law enforcement in preventing cybercrime

5.1 Types of Cybercrime and Their Impact on Society

Society has received many advantages from technological progress along with internet expansion. The digital environment now serves as a platform for new generations of criminals who use technology to carry out different types of cybercrimes. Different criminal activities that take place in the digital world lead to substantial harm for people and companies and all elements of society. The discussion focuses on multiple cybercrime types along with their significant social effects on society based on references [154, 219, 223, 202, 198].

1. Identity Theft: Identity theft represents one of the most destructive types of cybercrimes which currently affects the public. The crime consists of obtaining personal information without authorization followed by its unauthorized utilization of social security numbers and bank account details together with credit card information. Different types of cyberattacks including phishing and hacking and malware utilization serve cybercriminals to acquire valuable information. Identity theft victims face severe negative effects which result in destruction of their finances and credit problems together with psychological suffering.
2. Financial Fraud: Financial fraud within cyberspace includes multiple criminal activities which aim to obtain money through illegal means. The category includes fraudulent online schemes and phishing attacks and investment swindles and unauthorized credit card transactions. The financial vulnerabilities of people become targets for sophisticated deceptive strategies used by cybercriminals to achieve their goals. Financial fraud affects more than just individual victims because it damages both financial system stability and economic development while eroding public trust in these systems.
3. Cyberbullying and Online Harassment: The growing use of social media and digital communication tools has created conditions for both cyberbullying and online harassment to occur. Technology-based cybercrime includes using digital tools to threaten or humiliate or intimidate people. Through the internet's anonymity cyberbullies have increased opportunities to select and abuse their victims. Victims of cyberbullying frequently experience severe psychological and emotional problems which can trigger depression alongside anxiety and suicidal tendencies.
4. Data Breaches: Organizations face unauthorized data exposure when unauthorized persons successfully access their stored sensitive information. Organization data breaches occur from hacking attempts as well as from malware attacks and internal threats. The acquired data becomes valuable to cybercriminals who use it for identity theft or financial fraud and then sell the information

through dark web platforms. Organizations that fail to protect sensitive data experience both a violation of personal privacy and a loss of public trust in their data protection responsibilities.

5. Cyberterrorism: The definition of cyberterrorism includes digital attacks which disrupt essential infrastructure as well as public services and government operations. Through their attacks cyberterrorists aim to disrupt power grids and transportation systems and financial networks and communication infrastructure to create widespread disorder and public fear. A cyberterrorist attack would lead to extensive disruptions as well as economic losses and threats against national security.

Society feels the extensive consequences of cybercrime because of its various impacts. The theft of identities along with fraudulent activities and extortion tactics lead to substantial financial losses that affect both individuals and business operations. The losses create personal financial difficulties and economic instability and prevent organizations from prospering. Online transactions become less trustworthy while data security declines which leads to reduced e-commerce growth and delayed technological progress and digital innovation.

The psychological health of people suffers greatly because of cybercrime activities. People who experience cyberbullying and online harassment develop long-term emotional issues which cause their self-esteem to decrease and lead to social withdrawal and mental health problems. The apprehension about cybercrimes produces an environment of doubt that prevents people from adopting digital benefits to their fullest extent.

The legal systems together with law enforcement agencies across the world face important difficulties due to cybercrime. The anonymity of the internet combined with its borderless nature prevents law enforcement from effectively identifying and prosecuting cybercriminals. The feeling of no consequences enables criminals to carry out more crimes thus intensifying societal problems.

In conclusion, the various types of cybercrime threaten both individuals and organizations as well as all members of society. The various cybercrimes identified in this session demonstrate the necessity of strong cybersecurity protocols combined with public education and worldwide collaboration to fight this widespread threat. Society needs efforts to build stronger legal systems alongside better digital education and responsible online conduct to reduce the effects of cybercrime. A safer digital environment will emerge through combined efforts that enable trust and innovation along with internet potential.

5.2 Methods Used to Prevent and Detect Cybercrime

Safeguarding people and organizations and society at large in the current digital era depends on effective cybercrime prevention and detection measures. The attackers use state-of-the-art tactics to discover weaknesses in order to carry out their criminal activities on the web. The protection of cybercrime requires different methods and strategies for its prevention and detection. This session examines various strategies which fight cybercrime and preserve the security of digital space.

1. Organizations must implement robust cybersecurity measures to stop and identify cybercrime. Organizations must implement strong firewalls together with secure network configurations alongside periodic system and software updates to protect against known vulnerabilities. Security is enhanced by applying encryption protocols to store and send information. Organizations should implement strict access controls and strong authentication mechanisms and provide regular employee training to enhance cybersecurity awareness and implement best practices.
2. Risk assessments along with vulnerability management serve as essential measures for preventing cybercrime. Organizations need to discover system and network weaknesses and application vulnerabilities through risk assessment to take proactive measures for vulnerability mitigation. Organizations need to use security tools for vulnerability scanning while they patch software

quickly and maintain updated security protocols. Organizations that remain vigilant lower their chances of successful cyberattacks.

3. The prevention of cybercrime heavily depends on education and awareness programs. The implementation of training programs together with educational resources about online security dangers along with safe internet practices and password protection enables people to defend themselves against cyber threats. An organization-wide cybersecurity culture together with employee digital responsibility practices creates substantial improvements in security posture.
4. A properly developed incident response plan serves as the key component for both cybercrime detection and mitigation. The response plan should detail the immediate actions to take during a cyber incident followed by containment procedures and investigation steps and recovery procedures. A well-developed incident response plan should clearly outline the duties of staff members together with reporting and documentation procedures and communication protocols. Organizations should conduct regular tests and updates of their incident response plan to achieve efficient cyber incident response capabilities.
5. Organizational detection of cybercrime requires information sharing among different entities which include organizations and government agencies and law enforcement entities. Organizations that share threat intelligence and attack patterns together with best practices can better discover new threats and strengthen their combined response capabilities. The fight against cybercriminals benefits greatly from public-private partnerships combined with industry forums and international cooperation to achieve better collaboration.
6. Advanced analytics together with artificial intelligence (AI) technologies improve both cybercrime prevention and detection processes. AI algorithms process big data collections to find regularities while monitoring irregularities which signal possible cyber threats. The continuous adaptation of machine learning models to new threats enhances both detection precision and operational speed. The implementation of AI systems enables automated threat hunting as well as incident response and malware detection functions which speed up response times and minimize cybercrime damage.

Multiple approaches are needed to fight cybercrime which include strong security systems, awareness programs, incident response strategies and technology-based measures and collaboration between organizations. The implementation of these methods enables better protection against cyber threats for individuals and organizations and society while reducing cybercrime occurrences and minimizing its impact. The protection of cyberspace depends on ongoing cybersecurity research together with public-private partnerships and worldwide cooperation to outpace cybercriminals.

5.3 Notable Cybersecurity Breaches Put Focus on Vulnerabilities

Recent major data breaches across various sectors have demonstrated the cybersecurity risks that governments, businesses and individual users currently face. The expansion of digital systems creates new security risks which criminals together with hacktivists and hostile state actors can exploit. Current incidents prove that organizations need to enhance their security measures.

The 2013 breach at Yahoo revealed that all 3 billion user accounts contained names along with emails and passwords and other personal details. In 2017, credit monitoring agency Equifax was hacked, compromising the personal data of nearly 150 million people. The 2019 CapitalOne bank breach impacted 100 million customers. Colonial Pipeline faced a ransomware attack in 2021 which interrupted fuel delivery operations in the southeastern region of the United States.

The cybersecurity breaches which occurred in 2023 affected multiple sectors across different industries.

- A security event occurred at Infosys on November 5 when applications in its U.S. unit Infosys McCamish Systems faced ongoing investigation into the resulting effects.
- Boeing reported to the public on November 2 about its cyber incident which affected various business operations and received support from law enforcement regarding the case. The ransomware attack by LockBit ransomware gang remains unverified because there is no evidence that they received any ransom payment.
- The Indian Council of Medical Research became victim to a data breach which potentially revealed 815 million health records of Indian citizens while opposition parties demanded a government investigation into state data protection.
- Okta experienced a security incident on October 19 when its support case management system faced unauthorized access through stolen user credentials.
- The airline asked customers to cancel their credit cards following a security breach which exposed financial details.
- 23andMe experienced a credential stuffing attack on October 6 that resulted in the theft of customer genetic information from people with particular ancestral backgrounds.

Government agencies have also proven vulnerable. A system breach at the IRS in 2015 exposed more than 700,000 tax accounts. In 2016 Russian hackers suspected of conducting cyber attacks entered the Democratic National Committee network. The 2017 breach of credit monitoring firm Equifax exposed sensitive information belonging to hundreds of U.S. military and government personnel.

The attempted attacks on essential power infrastructure together with nuclear facilities reveal the progression of modern cyber dangers. The state-sponsored hacking operations become more complex with each passing year. The methods used by hackers now include ransomware and malware as well as phishing attacks and social engineering schemes for victimizing their targets.

The rising number of incidents reveals security weaknesses even though organizations continue to invest more in cybersecurity measures. The combination of weak controls with outdated systems and insufficient employee training makes it possible for breaches to occur because criminal hackers have become more sophisticated. The public and private sectors need proactive risk management together with data protections and access controls as well as penetration testing and emergency planning to defend against future cyber-attacks. Both public and private organizations need to maintain constant alertness while building resilience to protect against ongoing threats that continue to evolve.

5.4 Ethical Considerations Surrounding Cybersecurity

Ethics form a central component that influences all cybersecurity actions and organizational decisions. The continuous advancement of technology alongside increasingly complex cyber threats demands cybersecurity professionals to work with strong ethical principles. The presentation investigates cybersecurity ethical issues by examining privacy standards and disclosure procedures and security versus personal liberty regulations.

1. Protecting individual privacy along with personal data constitutes an essential ethical priority for cybersecurity. All organizations and individuals must respect privacy rights and follow established laws regarding personal information collection storage and usage. Organizations must establish a proper equilibrium between their cybersecurity practices and privacy rights since overextensive surveillance or data collection methods risk violating individual freedoms.
2. The essential ethical principles of transparency and accountability must be present in cybersecurity domains. Organizations must show transparency about their cybersecurity practices through

disclosure of their data protection methods and warning users about potential dangers and safety measures. Organizations have to demonstrate responsibility for safeguarding systems and data under their control because they must take responsibility for all security breaches and system vulnerabilities that happen.

3. The practice of responsible disclosure involves disclosing system or software weaknesses to the correct receiving parties in an ethical manner. White-hat hackers who function as ethical hackers serve as essential elements for vulnerability identification and reporting to assist organizations in enhancing their security defenses. Organizations must develop proper channels alongside disclosure procedures that allow vulnerabilities to receive prompt attention to protect both users and organizations.
4. The ethical challenge of cybersecurity requires organizations to determine appropriate measures for protecting security while respecting individual freedoms. Security measures need to be maintained at a proper level because strong cybersecurity requires protection of personal freedoms and civil liberties. Maintaining equilibrium between security and privacy demands thorough evaluation of legal frameworks and ethical standards alongside social impacts.
5. The problem of cybersecurity crosses international borders since it affects the whole world. Ethical considerations need to develop international collaboration to fight cyber threats. Different nations can improve their collective cybercrime defense through information and intelligence and best practices sharing which respects their sovereignty and cultural diversity. The development of universal cybersecurity norms and ethical standards leads to safer digital environments for every nation.
6. The deployment of offensive cyber capabilities necessitates ethical evaluation of their usage. Ethical frameworks should direct the deployment of such capabilities through responsible and proportional actions. The distinction between defensive measures for cyber protection and offensive operations that lead to unintended effects and cyber conflict escalation needs clear recognition.

In conclusion, ethical foundations stand as the essential base for all responsible cybersecurity practices. A secure ethical cyberspace requires protection of privacy together with transparent practices and responsible disclosure and security measures that respect personal freedoms and global collaboration and proper usage of offensive capabilities. Following ethical principles in cybersecurity practice maintains alignment between technological efforts and societal values and protects individual rights while building a safer digital world. Our ability to navigate cybersecurity complexities depends on upholding ethical standards because this approach promotes responsible actions and trust alongside accountability.

5.5 The Role of Law Enforcement in Preventing Cybercrime

The advancement of cybercrime represents a major threat during the digital era thus making the duties of law enforcement agencies essential for stopping and fighting these offenses. Cybercriminals use internet anonymity together with its interconnectedness to execute criminal activities thus law enforcement needs to work jointly and proactively. The seminar discusses the essential contribution of law enforcement agencies to cybercrime prevention as well as the difficulties they encounter within this ever-changing environment.

1. Cybercrime investigations are essential for law enforcement because they work to catch cybercriminals through their specialized units along with trained personnel who analyze digital evidence and track cyberattack origins and find perpetrators. The specialized cybercrime units within law enforcement agencies along with their trained staff use their technical capabilities to study digital evidence while tracing cyberattacks and identifying criminal suspects. Law enforcement executes thorough investigations to stop cybercriminals from operating and to deliver justice to victims.

2. Law enforcement agencies actively promote cybersecurity education through their efforts to make public and business entities and other stakeholders aware of cybersecurity issues. These agencies distribute training materials through workshops which teach people how to protect their digital resources and how to detect potential threats online. Through awareness campaigns law enforcement provides tools that help people and businesses defend themselves from cybercrimes.
3. Domestic and international law enforcement agencies work together through information sharing programs to effectively combat cybercrime. Through joint investigations and intelligence exchange programs along with practice sharing law enforcement stays informed about emerging cyber threats and criminal tactics. By working together law enforcement agencies strengthen their combined response to cybercrime while sharing information and gaining better capability to capture and prosecute cybercriminals in different countries.
4. The law enforcement agencies actively contribute to the creation of cybercrime-related legislative frameworks and policy frameworks. The agencies maintain constant communication with policymakers and legal experts as well as industry stakeholders to guarantee that existing laws and regulations stay current with advancing cyber threats. Through their advocacy work for strong laws and active policies law enforcement agencies establish conditions which support the detection and prosecution of cybercriminals.
5. Cybercriminals frequently operate beyond country borders thus requiring international cooperation for effective cybercrime investigations. Law enforcement organizations from different nations unite to distribute intelligence data and run joint investigations and extradite cybercriminals. Through mutual legal assistance treaties and international law enforcement networks countries establish formal cooperation channels which enhance global efforts against cybercrime.
6. Cybercrime prevention presents multiple obstacles to law enforcement agencies. The fast technological development alongside hidden communications through encryption and dark web anonymity complicates the process of identifying and apprehending cybercriminals. The different legal frameworks of various nations together with jurisdictional problems create problems for international collaboration. Law enforcement agencies need to develop their abilities by investing in specialized training and forming partnerships with private sector organizations and using emerging technologies such as artificial intelligence and data analytics.
7. Encryption Backdoors: Balancing Privacy and Security to Fight Cybercrime. The growing use of strong encryption in messaging apps and smartphones creates an intense debate about the need for special access systems for law enforcement. Strong encryption protects privacy and security but provides investigators with tools to investigate crimes through backdoors. The two sides of this debate present valid points which continue to form a persistent conversation.

The FBI alongside other law enforcement agencies maintains that backdoors provide essential security benefits for public protection. According to law enforcement agencies the increasing difficulty of accessing locked devices and encrypted communications creates obstacles to investigations into terrorism along with child exploitation cases and other major criminal offenses. Evidence gathering through restricted encryption access would benefit evidence collection when monitored by authorities.

Multiple technologists together with privacy advocates maintain that encryption backdoors are fundamentally unsafe because they create security risks. The creation of backdoors that remain accessible only to authorized parties is technically impossible according to privacy advocates. The implementation of encryption vulnerabilities enables criminals and both foreign adversaries and rogue insiders to exploit the system. Fundamental liberties require users to possess encryption that cannot be cracked. Middle ground solutions work to achieve equilibrium between conflicting demands. One potential solution involves restricting

backdoor access to criminal cases that face judicial authorization. The use of third party escrows to split decryption keys presents another method to decrease potential risks. The majority of experts doubt that usable backdoors and strong encryption can exist together as a single system.

Both positions in this debate have convincing arguments. The need requires continuous dialog between experts who will evaluate particular backdoor concepts by weighing their advantages against their safety considerations. The implementation of backdoors by law enforcement without agreement from other stakeholders proves to be an unwise decision. An absolute ban on all exceptional encrypted access mechanisms stands against societal needs for public safety yet allowing any form of such access mechanisms also fails to make sense. A balanced approach between civil liberties and investigative capabilities presents the most cautious direction for future development.

In conclusion, the fight against cybercrime heavily depends on law enforcement organizations in their efforts to stop and investigate cyber crimes. Through their investigative efforts, education initiatives, collaboration, and policy advocacy, law enforcement agencies contribute to creating a safer digital environment for individuals, organizations, and society at large. The protection of cybercrime prevention and cybercriminal prosecution requires law enforcement agencies to receive adequate resources together with expert knowledge and appropriate legal infrastructure for digital ecosystem trust building. Law enforcement agencies must use proactive methods combined with adaptability to stay one step ahead of cybercriminals while safeguarding public interests during the digital era.

5.6 Case Study 1

RESEARCH ARTICLE

Automatic detection of cyberbullying in social media text

Cynthia Van Hee[1☯]*, Gilles Jacobs[1☯], Chris Emmery[2], Bart Desmet[1], Els Lefever[1], Ben Verhoeven[2], Guy De Pauw[2], Walter Daelemans[2], Véronique Hoste[1]

1 Department of Translation, Interpreting and Communication - Faculty of Arts and Philosophy, Ghent University, Ghent, Belgium, 2 Department of Linguistics - Faculty of Arts, University of Antwerp, Antwerp, Belgium

☯ *These authors contributed equally to this work.*
* *cynthia.vanhee@ugent.be*

Abstract

While social media offer great communication opportunities, they also increase the vulnerability of young people to threatening situations online. Recent studies report that cyberbullying constitutes a growing problem among youngsters. Successful prevention depends on the adequate detection of potentially harmful messages and the information overload on the Web requires intelligent systems to identify potential risks automatically. The focus of this paper is on automatic cyberbullying detection in social media text by modelling posts written by bullies, victims, and bystanders of online bullying. We describe the collection and fine-grained annotation of a cyberbullying corpus for English and Dutch and perform a series of binary classification experiments to determine the feasibility of automatic cyberbullying detection. We make use of linear support vector machines exploiting a rich feature set and investigate which information sources contribute the most for the task. Experiments on a hold-out test set reveal promising results for the detection of cyberbullying-related posts. After optimisation of the hyperparameters, the classifier yields an F_1 score of 64% and 61% for English and Dutch respectively, and considerably outperforms baseline systems.

Introduction

Web 2.0 has had a substantial impact on communication and relationships in today's society. Children and teenagers go online more frequently, at younger ages, and in more diverse ways (e.g. smartphones, laptops and tablets). Although most of teenagers' Internet use is harmless and the benefits of digital communication are evident, the freedom and anonymity experienced online makes young people vulnerable with cyberbullying being one of the major threats [1, 2].

Bullying is not a new phenomenon and cyberbullying has manifested itself as soon as digital technologies have become primary communication tools. On the positive side, social media like blogs, social networking sites (e.g. Facebook), and instant messaging platforms (e.g. WhatsApp) make it possible to communicate with anyone and at any time. Moreover, they are a place where people engage in social interaction, offering the possibility to establish new relationships and maintain existing friendships [3, 4]. On the negative side however, social media increase the risk of children being confronted with threatening situations including grooming or sexually transgressive behaviour, signals of depression and suicidal thoughts, and cyberbullying. Users are reachable 24/7 and are often able to remain anonymous if desired: this makes social media a convenient way for bullies to target their victims outside the school yard.

https://doi.org/10.1371/journal.pone.0203794 Used with CC 4.0 license

With regard to cyberbullying, a number of national and international initiatives have been launched over the past few years to increase children's online safety. Examples include *KiVa* (http://www.kivaprogram.net/), a Finnish cyberbullying prevention programme, the '*Non au harcèlement*' campaign in France, Belgian governmental initiatives and helplines (e.g. *clicksafe.be, veiligonline.be, mediawijs.be*) that provide information about online safety, and so on.

In spite of these efforts, a lot of undesirable and hurtful content remains online. [2] analysed a body of quantitative research on cyberbullying and observed cybervictimisation rates among teenagers between 20% and 40%. [5] focused on 12 to 17 year olds living in the United States and found that no less than 72% of them had encountered cyberbullying at least once within the year preceding the questionnaire. [6] surveyed 9 to 26 year olds in the United States, Canada, the United Kingdom and Australia, and found that 29% of the respondents had ever been victimised online. A study among 2,000 Flemish secondary school students (age 12 to 18) revealed that 11% of them had been bullied online at least once in the six months preceding the survey [7]. Finally, the 2014 large-scale EU Kids Online Report [8] published that 20% of 11 to 16 year olds had been exposed to hate messages online. In addition, youngsters were 12% more likely to be exposed to cyberbullying as compared to 2010, which clearly demonstrates that cyberbullying is a growing problem.

The prevalence of cybervictimisation depends on the conceptualisation used in describing cyberbullying, but also on research variables such as location and the number and age span of the participants. Nevertheless, the above studies demonstrate that online platforms are increasingly used for bullying, which is a cause for concern given its impact. As shown by [9–11], cyberbullying may negatively impact the victim's self-esteem, academic achievement and emotional well-being. [12] found that self-reported effects of cyberbullying include negative effects on school grades and feelings of sadness, anger, fear, and depression. In extreme cases, cyberbullying could even lead to self-harm and suicidal thoughts.

These findings demonstrate that cyberbullying is a serious problem the consequences of which can be dramatic. Early detection of cyberbullying attempts is therefore of key importance to youngsters' mental well-being. Successful detection depends on effective monitoring of online content, but the amount of information on the Web makes it practically unfeasible for moderators to monitor all user-generated content manually. To tackle this problem, intelligent systems are required that process this information in a fast way and automatically signal potential threats. This way, moderators can respond quickly and prevent threatening situations from escalating. According to recent research, teenagers are generally in favour of such automatic monitoring, provided that effective follow-up strategies are formulated, and that privacy and autonomy are guaranteed [13].

Parental control tools (e.g. *NetNanny*, https://www.netnanny.com/) already block unsuited or undesirable content and some social networks make use of keyword-based moderation tools (i.e. using lists of profane and insulting words to flag harmful content). However, such approaches typically fail to detect implicit and subtle forms of cyberbullying in which no explicit vocabulary is used. This creates the need for intelligent and self-learning systems that go beyond keyword spotting and hence improve the recall of cyberbullying detection.

The ultimate goal of this type of research is to develop models that could improve manual monitoring for cyberbullying on social networks. We explore the automatic detection of textual signals of cyberbullying, in which cyberbulying is approached as a complex phenomenon that can be realised in various ways (see the Annotation guidelines section for a detailed overview). While the vast majority of the related research focuses on detecting cyberbullying 'attacks' (i.e. verbal aggression), the present study takes different types of cyberbullying into account, including more implicit posts from the bully, but also posts written by victims and bystanders. This is a more inclusive conceptualisation for the task of cyberbullying detection and should aid in moderation and prevention efforts by capturing different and more implicit signals of bullying.

https://doi.org/10.1371/journal.pone.0203794

To tackle this problem, we propose a machine learning method based on a linear SVM classifier [14, 15] exploiting a rich feature set. The contribution we make is twofold: first, we develop a complex classifier to detect *signals* of cyberbullying, which allows us to detect different types of cyberbullying that are related to different social roles involved in a cyberbullying event. Second, we demonstrate that the methodology is easily portable to other languages, provided there is annotated data available, by performing experiments on an English and Dutch dataset.

The remainder of this paper is structured as follows: the next section presents a definition of cyberbullying and its participant roles and provides an overview of the state of the art in cyberbullying detection. The *Data collection and annotation* section describes the corpus construction and annotation. Next, we present the experimental setup and discuss our experimental results for English and Dutch. Finally, the *Conclusion and future research* section concludes this paper and provides some perspectives for further research.

Related research

Both offline and online bullying are widely covered in the realm of social sciences and psychology, and the increasing number of cyberbullying cases in recent years [16] has stimulated research efforts to detect cyberbullying automatically. In the following section, we present a definition of cyberbullying and identify its participant roles and we provide a brief overview of automatic approaches to cyberbullying detection.

Cyberbullying definition and participant roles

A common starting point for conceptualising cyberbullying are definitions of traditional (i.e. *offline*) bullying, one of the most influential ones being formulated by [17]. The researcher described bullying based on three main criteria, including i) **intention** (i.e. a bully intends to inflict harm on the victim), ii) **repetition** (i.e. bullying acts take place repeatedly over time) and iii) a **power imbalance** between the bully and the victim (i.e. a more powerful bully attacks a less powerful victim). With respect to cyberbullying, a number of definitions are based on the above criteria. A popular definition is that of [18, p. 376], which describes cyberbullying as "an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time, against a victim who cannot easily defend him or herself". However, opinion on the applicability of the above characteristics to cyberbullying is very much divided [19], and besides theoretical objections, a number of practical limitations have been observed. Firstly, while [17] claims intention to be inherent to traditional bullying, this is much harder to ascertain in an online environment. Online conversations lack the signals of a face-to-face interaction like intonation, facial expressions and gestures, which makes them more ambiguous than real-life conversations. The receiver may therefore get the wrong impression that they are being offended or ridiculed [20]. Another criterion for bullying that might not hold in online situations is the power imbalance between the bully and the victim. This can be evident in real life (e.g. the bully is taller, stronger or older than the victim), but it is hard to conceptualise or measure online, where power may be related to technological skills, anonymity or the inability of the victim to escape from the bullying [19, 21]. Also empowering for the bully are inherent characteristics of the Web: once defamatory or confidential information is made public through the Internet, it is hard to remove.

Finally, while arguing that repetition distinguishes bullying from single acts of aggression, [17] himself states that such a single aggressive action can be considered bullying under certain circumstances. Accordingly, [21] claim that repetition in cyberbullying is problematic to operationalise, as it is unclear what the consequences are of a single derogatory message on a public page. A single act of aggression or humiliation may cause continued distress and humiliation for the victim if it is shared or liked by a large audience [21]. [22, p. 26] compare this with the "snowball effect": one post may be repeated or distributed by other people so that it becomes out of the control of the initial bully and has larger effects than was originally intended.

https://doi.org/10.1371/journal.pone.0203794

Given these arguments, a number of less 'strict' definitions of cyberbullying were proposed by among others [2, 5, 6], where a power imbalance and repetition are not deemed necessary conditions for cyberbullying.

The above paragraphs demonstrate that defining cyberbullying is far from trivial, and varying prevalence rates (see the Introduction section) confirm that a univocal definition of the phenomenon is still lacking in the literature [2]. Based on existing conceptualisations, we define cyberbullying as *content that is published online by an individual and that is aggressive or hurtful against a victim*. Based on this definition, an annotation scheme was developed [23] to signal textual characteristics of cyberbullying, including posts from bullies, as well as reactions from victims and bystanders.

Cyberbullying research also involves the identification of its participant roles. [24] were among the first to define the roles in a bullying situation. Based on surveys among teenagers involved in real-life bullying situations, they defined six participant roles: victims (i.e. who are the target of repeated harassment), bullies (i.e. who are the initiative-taking perpetrators), assistants of the bully (i.e. who encourage the bullying), reinforcers of the bully (i.e. who reinforce the bullying), defenders (i.e. who comfort the victim, take their side or try to stop the bullying) and outsiders (i.e. who ignore or distance themselves from the situation). In sum, in addition to the bully and victim, the researchers distinguish four bystanders (i.e. assistants, reinforcers, defenders and outsiders). [25], however, do not distinguish between reinforcers and assistants of the bullying. Their typology includes victims, bullies and three types of bystanders: i) bystanders who participate in the bullying, ii) bystanders who help or support the victim and iii) bystanders who ignore the bullying. The cyberbullying roles that are identified in our annotation scheme are based on existing bullying role typologies, given that traditional bullying roles are applicable to cyberbullying as well [26, 27]. More details about the different roles that we take into account are provided in the Data collection and annotation section.

Bystanders and -to a lesser extent- victims are often overlooked in the related research. As a result, these studies can be better characterised as verbal aggression detection concerned with retrieving bully attacks. By taking bystanders into account, we capture different and more subtle signals of a bullying episode. Note that while in this work we did not include classification of the participant roles as such, they are essential to the conceptualisation of the current detection task.

Detecting and preventing cyberbullying

As mentioned earlier, although research on cyberbullying detection is more limited than social studies on the phenomenon, some important advances have been made in recent years. In what follows, we present a brief overview of the most important natural language processing approaches to cyberbullying detection, but we refer to the survey paper by [28] for a more detailed overview.

Although some studies have investigated the effectiveness of rule-based modelling [29], the dominant approach to cyberbullying detection involves machine learning. Most machine learning approaches are based on supervised [30, 30–32] or semi-supervised learning [33]. The former involves the construction of a classifier based on labelled training data, whereas semi-supervised approaches rely on classifiers that are built from a training corpus containing a small set of labelled and a large set of unlabelled instances. Semi-supervised methods are often used to handle data sparsity, a typical issue in cyberbullying research. As cyberbullying detection essentially involves the distinction between bullying and non-bullying posts, the problem is generally approached as a binary classification task where the positive class is represented by instances containing (textual) cyberbullying, while the negative class is devoid of bullying signals.

A key challenge in cyberbullying research is the availability of suitable data, which is necessary to develop models that characterise cyberbullying. In recent years, only a few datasets have become publicly available for this particular task, such as the training sets provided in the context of the CAW 2.0 workshop

https://doi.org/10.1371/journal.pone.0203794

(http://caw2.barcelonamedia.org), a MySpace (https://myspace.com) [34] and Formspring (http://www.formspring.me) cyberbullying corpus annotated with the help of Mechanical Turk [29], and more recently, the Twitter Bullying Traces dataset [35]. Many studies have therefore constructed their own corpus from social media websites that are prone to bullying content, such as YouTube [30, 32], Twitter [36, 37], Instagram [38], MySpace [31, 34], FormSpring [29, 39], Kaggle [40] and ASKfm [41]. Despite the bottleneck of data availability, cyberbullying detection approaches have been successfully implemented over the past years and the relevance of automatic text analysis techniques to ensure child safety online has been recognised [42].

Among the first studies on cyberbullying detection are [29–31], who explored the predictive power of *n*-grams (with and without tf-idf weighting), part-of-speech information (e.g. first and second pronouns), and sentiment information based on (polarity and profanity) lexicons for this task. Similar features were not only exploited for coarse-grained cyberbullying detection, but also for the detection of more fine-grained cyberbullying categories [41]. Despite their apparent simplicity, content-based features (i.e. lexical, syntactic and sentiment information) are very often exploited in recent approaches to cyberbullying detection [33, 43]. In fact, as observed by [28], more than 41 papers have approached cyberbullying detection using content-based features, which confirms that this type of information is crucial for the task.

More and more, however, content-based features are combined with semantic features derived from topic model information [44], word embeddings and representation learning [43, 45]. More recent studies have also demonstrated the added value of user-based information for the task, more specifically by including users' activities (i.e. the number of posts) on a social network, their age, gender, location, number of friends and followers, and so on [32, 33, 46, 47]. A final feature type that gains increasing popularity in cyberbullying detection are network-based features, whose application is motivated by the frequent use of social media data for the task. By using network information, researchers aim to capture social relations between participants in a conversation (e.g. bully versus victim), and other relevant information such as the popularity of a person (i.e. which can indicate the power of a potential bully) on a social network, the number of (historical) interactions between two people, and so on. [48] for instance used network-based features to take the behavioural history of a potential bully into account. [49] detected cyberbullying in tweets and included network features inspired by Olweus' [17] bullying conditions (see supra). More specifically, they measured the power imbalance between a bully and victim, as well as the bully's popularity based on interaction graphs and the bully's position in the network.

As mentioned earlier, social media are a commonly used genre for this type of tasks. More recently, researchers have investigated cyberbullying detection in multi-modal data offered by specific platforms. For instance [38] explored cyberbullying detection using multi-modal data extracted from the social network Instagram. More precisely, they combined textual features derived from the posts themselves with user metadata and image features and showed that integrating the latter enhanced the classification performance. [37] also detected cyberbullying in different data genres, including ASKfm, Twitter, and Instagram. They took role information into account by integrating bully and victim scores as features, based on the occurrence of bully-related keywords in their sent or received posts.

With respect to the datasets used in cyberbullying research, it can be observed that corpora are often composed by keyword search (e.g. [43, 44]), which produces a biased dataset of positive (i.e. bullying) instances. To balance these corpora, negative data are often added from a background corpus or data resampling [50] techniques are adopted [33, 47]. For this research, data were randomly crawled across ASKfm and no keyword search was used to collect bullying data. Instead, all instances were manually annotated for the presence of bullying. As a result, our corpus contains a realistic distribution of bullying instances.

When looking at the performance of automatic cyberbullying, we see that scores vary greatly and do not only depend on the implemented algorithm and parameter settings, but also on a number of other variables.

https://doi.org/10.1371/journal.pone.0203794

These include the metrics that are used to evaluate the system (i.e. micro- or macro-averaged F_1, precision, recall, AUC, etc.), the corpus genre (i.e. Facebook, Twitter, ASKfm, Instagram) and class distribution (i.e. balanced or unbalanced), the annotation method (i.e. automatic annotations or manual annotations using crowdsourcing or by experts) and, perhaps the most important distinguishing factor, the conceptualisation of cyberbullying that is used. More concretely, while some approaches identify sensitive topics [30] or insulting language [29], others propose a more comprehensive approach by capturing different types of cyberbullying [41] or by modelling the bully-victim communications involved in a cyberbullying incident [37].

The studies discussed in this section demonstrated the variety of approaches that have been used to tackle cyberbullying detection. However, most of them focused on cyberbullying 'attacks', or posts written by a bully. Moreover, it is not entirely clear if different forms of cyberbullying were taken into account (e.g. sexual intimidation or harassment, or psychological threats), in addition to derogatory language or insults. In the present study, cyberbullying is considered a complex phenomenon comprising different forms of harmful online behaviour, which are described in more detail in our annotation scheme [23]. Purposing to facilitate manual monitoring efforts on social networks, we developed a system that automatically detects signals of cyberbullying, including attacks from bullies, as well as victim and bystander reactions, the latter of which are generally overlooked in related research.

Most similar to this research is the work by [44], [43, 45], who investigated bullying traces posted by different author roles (e.g. bully, victim, bystander, assistant, defender, reporter, accuser, reinforcer). However, they collected tweets using the keywords *bully, bullied* and *bullying*. As a result, their corpus contained many reports or testimonials of cyberbullying (example 1), instead of actual cyberbullying. Moreover, their method implies that cyberbullying signals that are devoid of such keywords are not included in the training corpus.

1. *"Some tweens got violent on the n train, the one boy got off after blows 2 the chest. . . Saw him cryin as he walkd away: (bullying not cool"* [44, *p. 658*]

What clearly distinguishes these works from the present is that their conceptualisation of cyberbullying is not explained. It is, in other words, not clear which type of posts are considered bullying and which are not. In the present research, we identify different types of bullying and all are included in the positive class of our experimental corpus.

For this research, English and Dutch social media data were annotated for fine-grained forms of cyberbullying, based on the actors involved in a cyberbullying incident. After preliminary experiments for Dutch [41, 51], we currently present an optimised cyberbullying detection method for English and Dutch and hereby show that the proposed methodology can easily be applied to different languages, provided that annotated data are available.

Data collection and annotation

To be able to build representative models for cyberbullying, a suitable dataset is required. This section describes the construction of two corpora, English and Dutch, containing social media posts that are manually annotated for cyberbullying according to our fine-grained annotation scheme. This allows us to cover different forms and participants (or *roles*) involved in a cyberbullying event.

Data collection

Two corpora were constructed by collecting data from the social networking site ASKfm, where users can create profiles and ask or answer questions, with the option of doing so anonymously. ASKfm data typically consists of question-answer pairs published on a user's profile. The data were retrieved by crawling a number

https://doi.org/10.1371/journal.pone.0203794

of seed profiles using the GNU Wget software (http://www.gnu.org/software/wget/) in April and October, 2013. After language filtering (i.e. non-English or non-Dutch content was removed), the experimental corpora comprised 113,698 and 78,387 posts for English and Dutch, respectively.

Data annotation

Cyberbullying has been a widely covered research topic recently and studies have shed light on direct and indirect types of cyberbullying, implicit and explicit forms, verbal and non-verbal cyberbullying, and so on. This is important from a sociolinguistic point of view, but knowing what cyberbullying involves is also crucial to build models for automatic cyberbullying detection. In the following paragraphs, we present our data annotation guidelines [23] and focus on different types and roles related to the phenomenon.

Types of cyberbullying

Cyberbullying research is mainly centered around the conceptualisation, occurrence and prevention of the phenomenon [1, 52, 53]. Sociolinguistic studies have identified different types of cyberbullying [12, 54, 55] and compared these types with forms of traditional or offline bullying [20]. Like traditional bullying, direct and indirect forms of cyberbullying have been identified. Direct cyberbullying refers to actions in which the victim is directly involved (e.g. sending a virus-infected file, excluding someone from an online group, insulting and threatening), whereas indirect cyberbullying can take place without awareness of the victim (e.g. *outing* or publishing confidential information, spreading gossip, creating a hate page on social networking sites) [20].

The present annotation scheme describes some specific textual categories related to cyberbullying, including threats, insults, defensive statements from a victim, encouragements to the harasser, etc. (see the Data collection and annotation section for a complete overview). All of these forms were inspired by social studies on cyberbullying [7, 20] and manual inspection of cyberbullying examples.

Roles in cyberbullying

Similarly to traditional bullying, cyberbullying involves a number of participants that adopt well-defined roles. Researchers have identified several roles in (cyber)bullying interactions. Although traditional studies on bullying have mainly concentrated on bullies and victims [24], the importance of bystanders in a bullying episode has been acknowledged [56, 57]. Bystanders can support the victim and mitigate the negative effects caused by the bullying [57], especially on social networking sites, where they hold higher intentions to help the victim than in real life conversations [58]. [25] distinguish three main types of bystanders: i) bystanders who participate in the bullying, ii) who help or support the victim and iii) those who ignore the bullying. Given that passive bystanders are hard to recognise in online text, only the former two are included in our annotation scheme.

Annotation guidelines

To operationalise the task of automatic cyberbullying detection, we elaborated a detailed annotation scheme for cyberbullying that is strongly embedded in the literature and applied it to our corpora. The applicability of the scheme was iteratively tested. Our final guidelines for the fine-grained annotation of cyberbullying are described in a technical report [23]. The objective of the scheme was to indicate several types of textual cyberbullying and verbal aggression, their severity, and the author participant roles. The scheme is formulated to be generic and is not limited to a specific social media platform. All messages were annotated in context (i.e. presented within their original content or conversation event) when available.

https://doi.org/10.1371/journal.pone.0203794

Essentially, the annotation scheme describes two levels of annotation. Firstly, the annotators were asked to indicate, at the message or post level, whether the text under investigation was related to cyberbullying. If the message was considered harmful and thus contained indications of cyberbullying, annotators identified the author's participant role. Based on the literature on role-allocation in cyberbullying episodes [25, 59], four roles are distinguished in the annotation scheme, including victim, bully, and two types of bystanders.

1. **Harasser or bully:** person who initiates the bullying.
2. **Victim:** person who is harassed.
3. **Bystander-defender:** person who helps the victim and discourages the harasser from continuing his actions.
4. **Bystander-assistant:** person who does not initiate, but helps or encourages the harasser.

Secondly, at the sub-sentence level, the annotators were tasked with the identification of fine-grained text categories related to cyberbullying. In the literature, different forms of cyberbullying are identified [12, 54, 55] and compared with traditional bullying [20]. Based on these forms, the annotation scheme describes a number of textual categories that are often inherent to a cyberbullying event, such as threats, insults, defensive statements from a victim, encouragements to the harasser, etc. Most of the categories are related to direct forms of cyberbullying (as defined by [25]), while one is related to *outing* [25], an indirect form of cyberbullying, namely *defamation*. Additionally, a number of subcategories were defined to make the annotation scheme as concrete and distinctive as possible (e.g., *discrimination* as a subcategory of *insult*). All cyberbullying-related categories in the scheme are listed below, and an example post for each category is presented in Table 5.1.

Table 5.1 Definitions and brat annotation examples of more fine-grained text categories related to cyberbullying.

Annotation category	Annotation example
Threat/blackmail	[I am going to find out who you are & I swear you are going to regret it.]^{THREAT}
Insult	[Kill yourself]^{CURS} [you fucking mc slut!!!!]$^{\text{GEN. INSULT}}$ [NO ONE LIKES YOU!!!!!]$^{\text{GEN. INSULT}}$ [You are an ugly useless little whore!!!!]$^{\text{GEN. INSULT}}$
Curse/Exclusion	[Fuck you.]$^{\text{GEN. INSULT}}$ [Now shush I don't wanna hear anything.]$^{\text{CURSE OR EXCLUSION}}$
Defamation	[She slept with her ex behind his girlfiends back and she and him had broken up.]^{DEFAMATION}
Sexual Talk	[Naked pic of you now.]$^{\text{SEXUAL HARASSMENT}}$
Defense	[I would appreciate if you dindn't talk shit about my bestfriend.]$^{\text{GEN. VICTIM DEFENSE}}$ He has enough to deal with already.
Encour. to har.	[She is a massive slut]$^{\text{GEN. INSULT}}$ [i agree with you @user she is!]$^{\text{ENCOUR. HARASSER}}$ [LOL AT HER mate, im on your side]$^{\text{ENCOUR. HARASSER}}$

https://doi.org/10.1371/journal.pone.0203794.t001

- **Threat/blackmail:** expressions containing physical or psychological threats or indications of blackmail.
- **Insult:** expressions meant to hurt or offend the victim.
  - **General insult:** general expressions containing abusive, degrading or offensive language that are meant to insult the addressee.
  - **Attacking relatives:** insulting expressions towards relatives or friends of the victim.

    - **Discrimination:** expressions of unjust or prejudicial treatment of the victim. Two types of discrimination are distinguished (i.e. sexism and racism). Other forms of discrimination should be categorised as general insults.
- **Curse/exclusion:** expressions of a wish that some form of adversity or misfortune will befall the victim and expressions that exclude the victim from a conversation or a social group.
- **Defamation:** expressions that reveal confident or defamatory information about the victim to a large public.
- **Sexual Talk:** expressions with a sexual meaning or connotation. A distinction is made between innocent sexual talk and sexual harassment.
- **Defense:** expressions in support of the victim, expressed by the victim himself or by a bystander.
    - **Bystander defense:** expressions by which a bystander shows support for the victim or discourages the harasser from continuing his actions.
    - **Victim defense:** assertive or powerless reactions from the victim.
- **Encouragement to the harasser:** expressions in support of the harasser.
- **Other:** expressions that contain any other form of cyberbullying-related behaviour than the ones described here.

It is important to note that the categories were always indicated in text, even if the post in which they occurred was not considered harmful, for instance in the post "hi bitches, in for a movie?", "bitches" was annotated as an insult while the post itself was not considered cyberbullying.

To provide the annotators with some context, all posts were presented within their original conversation when possible. All annotations were done using the brat rapid annotation tool [60], some examples of which are presented in Table 5.1.

As can be deduced from the examples in the table, there were no restrictions as to what form the annotations could take. They could be adjectives, noun phrases, verb phrases, and so on. The only condition was that the annotation could not span more than one sentence and less than one word. Posts that were (primarily) written in another language than the corpus language (i.e. Dutch and English) were marked as such and required no further annotations.

We examined the validity of our guidelines and the annotations with an inter-annotator agreement experiment that is described in the following section.

Annotation statistics

The English and Dutch corpora were independently annotated for cyberbullying by trained linguists. All were Dutch native speakers and English second-language speakers. To demonstrate the validity of our guidelines, inter-annotator agreement scores were calculated using Kappa on a subset of each corpus. Inter-rater agreement for Dutch (2 raters) is calculated using Cohen's Kappa [61]. Fleiss' Kappa [62] is used for the English corpus (> 2 raters). Kappa scores for the identification of cyberbullying are $\kappa = 0.69$ (Dutch) and $\kappa = 0.59$ (English).

As shown in Table 5.2, inter-annotator agreement for the identification of the more fine-grained categories for English varies from fair to substantial [63], except for *defamation*, which appears to be more difficult to recognise. No encouragements to the harasser were present in this subset of the corpus. For Dutch, the inter-annotator agreement is fair to substantial, except for *curse* and *defamation*. Analysis revealed that one of both annotators often annotated the latter as an insult, and in some cases even did not consider it as cyberbullying-related.

https://doi.org/10.1371/journal.pone.0203794

Table 5.2 Inter-annotator agreement on the fine-grained categories related to cyberbullying.

	Threat	Insult	Defense	Sexual talk	Curse/ exclusion	Defamation	Encouragements to the harasser
English	0.65	0.63	0.45	0.38	0.58	0.15	N/A
Dutch	0.52	0.66	0.63	0.53	0.19	0.00	0.21

https://doi.org/10.1371/journal.pone.0203794.t002

In short, the inter-rater reliability study shows that the annotation of cyberbullying is not trivial and that more fine-grained categories like *defamation, curse* and *encouragements* are sometimes hard to recognise. It appears that defamations were sometimes hard to distinguish from insults, whereas curses and exclusions were sometimes considered insults or threats. The analysis further reveals that encouragements to the harasser are subject to interpretation. Some are straightforward (e.g. "I agree we should send her hate"), whereas others are subject to the annotator's judgment and interpretation (e.g. "hahaha", "LOL").

Experimental setup

In this paper, we explore the feasibility of automatically recognising signals of cyberbullying. A crucial difference with related research is that we do not only model bully 'attacks', but also more implicit forms of cyberbullying and reactions from victims and bystanders (i.e. all under one binary label 'signals of cyberbullying'), since these could likewise indicate that cyberbullying is going on. The experiments described in this paper focus on the automatic detection of such cyberbullying signals that need to be further investigated by human moderators when applied in a real-life moderation loop.

The English and Dutch corpus contain 113,698 and 78,387 posts, respectively. As shown in Table 5.3, the experimental corpus features a heavily **imbalanced** class distribution with the large majority of posts not being part of cyberbullying. In classification, this class imbalance can lead to decreased performance. We apply cost-sensitive SVM as a possible hyperparameter in optimisation to counter this. The cost-sensitive SVM reweighs the penalty parameter C of the error term by the inverse class-ratio. This means that misclassifications of the minority positive class are penalised more than classification errors on the majority negative class. Other pre-processing methods to handle data imbalance in classification include feature filtering metrics and data resampling [64]. These methods were omitted as they were found to be too computationally expensive given our high-dimensional dataset.

Table 5.3 Statistics of the English and Dutch cyberbullying corpus.

	Corpus size	Number(ratio) of bullying posts
English	113,698	5,375(4.73%)
Dutch	78,387	5,106(6.97%)

https://doi.org/10.1371/journal.pone.0203794.t003

For the automatic detection of cyberbullying, we performed **binary classification experiments** using a linear kernel support vector machine (SVM) implemented in LIBLINEAR [65] by making use of Scikit-learn [66], a machine learning library for Python. The motivation behind this is twofold: i) support vector machines (SVMs) have proven to work well for tasks similar to the ones under investigation [67] and ii) LIBLINEAR allows fast training of large-scale data which allow for a linear mapping (which was confirmed after a series of preliminary experiments using LIBSVM with linear, RBF and polynomial kernels).

The classifier was optimised for feature type (see the Pre-processing and feature engineering section) and hyperparameter combinations (see Table 5.4). Model selection was done using 10-fold cross validation in

grid search over all possible feature types (i.e. groups of similar features, like different orders of *n*-gram bag-of-words features) and hyperparameter configurations. The best performing hyperparameters are selected by F_1 score on the positive class. The winning model is then retrained on all held-in data and subsequently tested on a hold-out test set to assess whether the classifier is over- or under-fitting. The hold-out set represents a random sample (10%) of all data. The folds were randomly stratified splits over the hold-in class distribution. Testing all feature type combinations is a rudimentary form of feature selection and provides insight into which types of features work best for this particular task.

Table 5.4 Hyperparameters in grid-search model selection.

Hyperparameter	Values
Penalty of error term *C*	$1e^{(-3,-2,\ldots,2,3)}$
Loss function	Hinge, squared hinge
Penalty: norm used in penalisation	'11' ('least absolute deviations') or '12' ('least squares')
Class weight (sets penalty *C* of class *i* to weight* *C*)	None or 'balanced', i.e. weight inversely proportional to class frequencies

https://doi.org/10.1371/journal.pone.0203794.t004

Feature selection over all individual features was not performed because of the large feature space (NL: 795,072 and EN: 871,296 individual features). [68], among other researchers, demonstrated the importance of joint optimisation, where feature selection and hyperparameter optimisation are performed simultaneously, since the techniques mutually influence each other.

The optimised models are evaluated against **two baseline systems**: i) an unoptimised linear-kernel SVM (configured with default parameter settings) based on word *n*-grams only and, ii) a keyword-based system that marks posts as positive for cyberbullying if they contain a word from existing vocabulary lists composed by aggressive language and profanity terms.

Pre-processing and feature engineering

As pre-processing, we applied tokenisation, PoS-tagging and lemmatisation to the data using the LeTs Preprocess Toolkit [69]. In supervised learning, a machine learning algorithm takes a set of training instances (of which the label is known) and seeks to build a model that generates a desired prediction for an unseen instance. To enable the model construction, all instances are represented as a vector of features (i.e. inherent characteristics of the data) that contain information that is potentially useful to distinguish cyberbullying from non-cyberbullying content.

We experimentally tested whether cyberbullying events can be recognised automatically by lexical markers in a post. To this end, all posts were represented by a number of information sources (or *features*) including lexical features like bags-of-words, sentiment lexicon features and topic model features, which are described in more detail below. Prior to feature extraction, some data cleaning steps were executed, such as the replacement of hyperlinks and @-replies, removal of superfluous white spaces, and the replacement of abbreviations by their full form (based on an existing mapping dictionary: http://www.chatslang.com/terms/abbreviations/). Additionally, tokenisation was applied before *n*-gram extraction and sentiment lexicon matching, and stemming was applied prior to extracting topic model features.

After pre-processing of the corpus, the following feature types were extracted:

- **Word n-gram bag-of-words:** binary features indicating the presence of word unigrams, bigrams and trigrams.

- **Character n-gram bag-of-words:** binary features indicating the presence of character bigrams, trigrams and fourgrams (without crossing word boundaries). Character *n*-grams provide some abstraction from the word level and provide robustness to the spelling variation that characterises social media data.
- **Term lists:** one binary feature derived for each one out of six lists, indicating the presence of an item from the list in a post:
  - proper names: a gazetteer of named entities collected from several resources (e.g. Wikipedia).
  - 'allness' indicators (e.g. "always", "everybody"): forms which indicate rhetorical superlativity [70] which can be helpful in identifying the often hyperbolic bullying language.
  - diminishers (e.g. "slightly", "relatively"): diminishers, intensifiers and negation words were all obtained from an English grammar describing these lexical classes and existing sentiment lexicons (see further).
  - intensifiers (e.g. "absolutely", "amazingly")
  - negation words
  - aggressive language and profanity words: for English, we used the Google Profanity list (https://code.google.com/archive/p/badwordslist/downloads). For Dutch, a public profanity lexicon was consulted (http://scheldwoorden.goedbegin.nl).
    Person alternation is a binary feature indicating whether the combination of a first and second person pronoun occurs in order to capture interpersonal intent.
- **Subjectivity lexicon features:** positive and negative opinion word ratios, as well as the overall post polarity were calculated using existing sentiment lexicons. For Dutch, we made use of the Duoman [71] and Pattern [72] lexicons. For English, we included the Liu and Hu opinion lexicon [73], the MPQA lexicon [74], the General Inquirer Sentiment Lexicon [75], AFINN [76], and MSOL [77]. For both languages, we included the relative frequency of all 68 psychometric categories in the Linguistic Inquiry and Word Count (LIWC) dictionary for English [78] and Dutch [79].
- **Topic model features:** by making use of the Gensim topic modelling library [80], several LDA [81] and LSI [82] topic models with varying granularity ($k = 20$, 50, 100 and 200) were trained on data corresponding to each fine-grained category of a cyberbullying event (e.g. threats, defamations, insults, defenses). The topic models were based on a background corpus (EN: ± 1,200,000 tokens, NL: ± 1,400,000 tokens) scraped with the BootCaT [83] web-corpus toolkit. BootCaT collected ASKfm user profiles using lists of manually determined seed words that are characteristic of the cyberbullying categories.

When applied to the training data, this resulted in 871,296 and 795,072 features for English and Dutch, respectively.

Results

In this section, we present the results of our experiments to automatically detect cyberbullying signals in an English and Dutch corpus of ASKfm posts. Ten-fold cross-validation was performed in exhaustive grid search over different feature type and hyperparameter combinations (see the Experimental setup section). The *unoptimised word n-gram-based* classifier and *keyword-matching* system serve as baselines for comparison. Precision, Recall and F_1 performance metrics were calculated on the positive class. We also report Area Under the Receiver Operator Curve (AUROC) scores, a performance metric that is more robust to data imbalance than precision, recall and F score [84].

https://doi.org/10.1371/journal.pone.0203794

Table 5.5 gives us an indication of which feature type combinations score best and hence contribute most to this task. It presents the cross-validation and hold-out scores of a set of feature combinations, which are explained in the feature groups legend (Table 5.6). A total of 31 feature type combinations, each with 28 different hyperparameter sets have been tested. Table 5.5 shows the results for the three best scoring systems by included feature types with optimised hyperparameters. The maximum obtained F_1 score in cross-validation is 64.26% for English and 61.20% for Dutch and shows that the classifier benefits from a variety of feature types. The results on the hold-out test set show that the trained systems generalise well on unseen data, indicating little under- or overfitting. The simple keyword-matching baseline system has the lowest performance for both languages even though it obtains high recall for both languages, especially for English (80.14%), suggesting that profane language characterises many cyberbullying-related posts. Feature group and hyperparameter optimisation provides a considerable performance increase over the unoptimised word *n*-gram baseline system. The top-scoring systems for each language do not differ a lot in performance, except the best system for Dutch, which trades recall for precision when compared to the runner-ups.

Table 5.5 Cross-validated and hold-out scores (%) according to different metrics (F_1, precision, recall, accuracy and area under the curve) for the English and Dutch three best and worst combined feature type systems.

	Feature combination	Cross-validation scores					Hold-out scores				
		F_1	P	R	Acc	AUROC	F_1	P	R	Acc	AUROC
English											
Best three	B + C + D + E	**64.26**	73.32	57.19	96.97	78.07	63.69	74.13	55.82	97.21	77.47
	A + B + C	64.24	73.22	57.23	96.96	78.09	**64.32**	74.08	56.83	97.24	77.96
	A + C + E	63.84	73.21	56.59	96.94	77.78	62.94	72.82	55.42	97.14	77.24
Worst three	D	40.48	38.98	42.12	94.10	69.41	39.56	39.56	39.56	94.71	68.39
	A + D + E	38.95	31.47	51.10	92.37	72.76	40.71	33.87	51.00	93.49	73.22
	E	17.35	9.73	79.91	63.72	71.41	15.70	8.72	78.51	63.07	70.44
Baseline	word *n*-gram	58.17	67.55	51.07	96.54	74.93	59.63	69.57	52.17	96.57	75.50
	profanity	17.17	9.61	80.14	63.73	71.53	17.61	9.90	78.51	63.79	71.34
Dutch											
Best three	A + B + C + E	**61.20**	56.76	66.40	94.47	81.42	58.13	54.03	62.90	94.58	79.75
	A + B + C + D + E	61.03	71.55	53.20	95.53	75.86	**58.72**	67.40	52.03	95.62	75.21
	A + C + E	60.82	71.66	52.84	95.53	75.68	58.15	67.71	50.96	95.61	74.71
Worst three	D + B	32.90	29.23	37.63	89.91	65.61	30.16	34.72	26.65	92.61	61.73
	D	28.65	19.36	55.10	81.97	69.48	25.13	16.73	50.53	81.99	67.26
	B	24.74	21.24	29.61	88.16	60.94	17.99	23.15	14.71	91.98	55.80
Baseline	word *n*-gram	50.39	67.80	40.09	94.81	69.38	49.54	64.29	40.30	95.09	69.44
	profanity	28.46	19.24	54.66	81.99	69.28	25.13	16.73	50.53	81.99	67.26

https://doi.org/10.1371/journal.pone.0203794.t005

Table 5.6 Feature group mapping (Table 5.5).

A	word *n*-grams
B	subjectivity lexicons
C	character *n*-grams
D	term lists
E	topic models

https://doi.org/10.1371/journal.pone.0203794.t006

Table 5.7 presents the scores of the (hyperparameter-optimised) single feature type systems, to gain insight into the performance of these feature types when used individually. Analysis of the combined and single feature type sets reveals that **word *n*-grams, character *n*-grams**, and **subjectivity lexicons** prove to be strong features for this task. In effect, adding character *n*-grams always improved classification performance for both languages. They are likely to provide robustness to lexical variation in social media text, as compared to word *n*-grams. While subjectivity lexicons appear to be discriminative features, term lists perform badly on their own as well as in combinations for both languages. This shows once again (see the profanity baseline) that cyberbullying detection requires more sophisticated information sources than profanity lists. Topic models seem to do badly for both languages on their own, but in combination with other features, they improve Dutch performance consistently. A possible explanation for their varying performance in both languages would be that the topic models trained on the Dutch background corpus are of better quality than the English ones. In effect, a random selection of background corpus texts reveals that the English scrape contains more noisy data (i.e. low word-count posts and non-English posts) compared to the Dutch scraped corpus.

Table 5.7 Cross-validated and hold-out scores (%) according to different metrics (F_1, precision, recall, accuracy and area under the ROC curve) for English and Dutch single feature type systems.

	Feature type	Cross-validation scores					hold-out scores				
		F_1	P	R	Acc	AUROC	F_1	P	R	Acc	AUROC
English											
	word *n*-grams	**60.09**	60.49	59.69	96.22	78.87	**58.35**	57.12	59.64	96.27	78.79
	subjectivity lexicons	56.82	73.32	46.38	96.64	72.77	56.16	72.61	45.78	96.87	72.50
	character *n*-grams	52.69	58.70	47.80	95.91	73.06	53.33	62.37	46.59	96.43	72.65
	term lists	40.48	38.98	42.12	94.10	69.41	39.56	39.56	39.56	94.71	68.39
	topic models	17.35	9.73	79.91	63.72	71.41	15.70	8.72	78.51	63.07	70.44
Dutch											
	word *n*-grams	55.53	72.64	44.94	95.27	71.88	**54.99**	70.20	45.20	95.57	71.99
	subjectivity lexicons	54.34	54.12	54.56	93.97	75.65	51.82	50.61	53.09	94.09	74.90
	character *n*-grams	51.70	67.58	41.86	94.86	70.22	50.46	65.20	41.15	95.17	69.88
	term lists	28.65	19.36	55.10	81.97	69.48	25.13	16.73	50.53	81.99	67.26
	topic models	24.74	21.24	29.61	88.16	60.94	17.99	23.15	14.71	91.98	55.80

https://doi.org/10.1371/journal.pone.0203794.t007

A shallow **qualitative analysis** of the classification output provided insight into some of the classification mistakes.

Table 5.8 gives an overview of the error rates per cyberbullying category of the best performing and baseline systems. This could give an indication of the types of bullying are hard to detect by the current classifier. All categories are always considered positive for cyberbullying (i.e. the error rate equals the false negative rate), except for *Sexual* and *Insult* which can also be negative (in case of harmless sexual talk and 'socially acceptable' insulting language like "hi bitches, in for a movie?" the corresponding category was indicated, but the post itself was not annotated as cyberbullying) and *Not cyberbullying*, which is always negative. Error rates often being lowest for the profanity baseline confirms that it performs particularly well in terms of recall (at the expense of precision, see Table 5.5). When looking at the best system for both languages, we see that *Defense* is the hardest category to classify. This should not be a surprise as the category comprises defensive posts from bystanders and victims, which contain less aggressive language than cyberbullying attacks and are often shorter in length than the latter. Assertive defensive posts (i.e. a subcategory of *Defense*) which attack the bully are, however, more often correctly classified. There are not sufficient instances of the *Encouragement* class for either language in the hold-out set to be representative. In both languages, threats, curses and incidences of sexual harassment are most easily recognisable, showing (far) lower error rates than the categories *Defamation, Defense, Encouragements to the harasser*, and *Insult*.

Table 5.8 Error rates (%) per cyberbullying subcategory on hold-out for English and Dutch systems.

	Category	Nr. occurrences in hold-out	Profanity baseline	Word *n*-gram baseline	Best system
English					
	Curse	$n = 109$	14.68	30.28	24.77
	Defamation	$n = 21$	23.81	47.62	38.10
	Defense	$n = 165$	22.42	52.12	43.64
	Encouragement	$n = 1$	0.00	100.00	100.00
	Insult	$n = 345$	26.67	41.74	35.94
	Sexual	$n = 165$	63.80	21.47	21,47
	Threat	$n = 12$	8.33	41.67	25.00
	Not cyberbullying	$n = 10,714$	36.94	1.10	0.76
Dutch					
	Curse	$n = 96$	39.58	50.00	22.92
	Defamation	$n = 6$	100.00	66.67	33.33
	Defense	$n = 200$	52.50	63.50	46.00
	Encouragement	$n = 5$	40.00	60.00	40.00
	Insult	$n = 355$	43.38	47.89	28.17
	Sexual	$n = 37$	37.84	21.62	27.03
	Threat	$n = 15$	33.33	46.67	20.00
	Not cyberbullying	$n = 7,295$	15.63	1.23	3.07

https://doi.org/10.1371/journal.pone.0203794.t008

A qualitative error analysis of the English and Dutch predictions reveals that false positives often contain aggressive language directed at a second person, often denoting personal flaws or containing sexual and profanity words. We see that misclassifications are often short posts containing just a few words and that false negatives often lack explicit verbal signs of cyberbullying (e.g. insulting or profane words) or are ironic (examples 2 and 3). Additionally, we see that cyberbullying posts containing misspellings or grammatical errors and incomplete words are also hard to recognise as such (examples 4 and 5). The Dutch and English corpus are overall similar with respect to qualitative properties of classification errors.

2. *You might want to do some sports ahah x*
3. *Look who is there… my thousandth anonymous hater, congratulations*!
4. *ivegot 1 word foryou… yknow whatit is?* → *slut*
5. *One word for you*: *G—A—…*

In short, the experiments show that our classifier clearly outperforms both a keyword-based and word *n*-gram baseline. However, analysis of the classifier output reveals that false negatives often lack explicit clues that cyberbullying is going on, indicating that our system might benefit from irony recognition and integrating world knowledge to capture such implicit realisations of cyberbullying.

Our annotation scheme allowed to indicate different author roles, which provides better insight into the realisation of cyberbullying. Table 5.9 presents the error rates of our classifier for the different author roles, being harasser, victim, and two types of bystanders. We observe that the error rates are high for *bystander assistant* and *victim*, but there are not sufficient instances in the hold-out set of the former role for either

Table 5.9 Error rates (%) per cyberbullying participant role on hold-out for English and Dutch systems.

	Participant role	Nr. occurrences in hold-out	Profanity baseline	Word n-gram baseline	Best system
English					
	Harasser	$n = 328$	20.43	48.48	43.60
	Bystander assistant	$n = 2$	50.00	100.00	100.00
	Bystander defender	$n = 39$	7.69	38.46	25.64
	Victim	$n = 129$	27.91	57.36	50.39
	Not cyberbullying	$n = 10872$	37.64	1.24	0.89
Dutch					
	Harasser	$n = 261$	47.13	56.70	29.89
	Bystander assistant	$n = 6$	50.00	66.67	50.00
	Bystander defender	$n = 52$	25.00	38.46	23.08
	Victim	$n = 150$	62.00	72.00	54.00
	Not cyberbullying	$n = 7370$	16.01	1.42	3.41

https://doi.org/10.1371/journal.pone.0203794.t009

language to be representative. Error rates for the *victim* class of 50.39% and 54% in English and Dutch respectively indicate that the role is hard to recognise by the classifier. A possible explanation for this could be that victim posts in our corpus either expressed powerlessness facing the bully (example 6) or either contained explicit aggressive language as well (example 7).

6. *Your the one going round saying im a cunt and a twat and im ugly. tbh all im doing is sticking up for myself.*
7. *You're fucked up saying I smell from sweat, because unlike some other people I shower every day BITCH*

According to the figures, the most straightforward roles in detection are *bystander defender* and *harasser*.

In the light of comparison with state-of-the-art approaches to cyberbullying, we observe that competitive results are obtained with regard to [30–32, 41]. However, the fundamental differences with respect to data collection, sources, and conceptualisations of bullying hardly allow for direct comparison. Table 5.10 presents the experimental results obtained by [43–45] who, like the current study, approach the task as detecting posts from bullies as well as from victims and bystanders. Given their experimental setup (i.e. task description, data genre and classifier), their work can be considered most similar to ours so their results might function as benchmarks. Also here, a number of crucial differences with the current approach can be observed: Firstly, their corpora were collected using the keywords "bully", "bullying" and "bullied", which may bias the dataset towards the positive class and ensures that many explicit lexicalisations are present in the positive class. Second, it is not clear which types of cyberbullying (i.e. explicit and implicit bullying, threats, insults, sexual harassment) are included in the positive class. Furthermore, as can be deduced from Table 5.10, the datasets are considerably smaller than ours and show a more balanced class distribution (respectively 39% cyberbullying posts in [43] and [44], and 29%/26% in [45]) than the ratio of bullying posts in our corpus (see Table 5.3: 5% for English, 7% for Dutch). Hence, any comparison should be made with caution due to these differences.

Table 5.10 Overview of the most related cyberbullying detection approaches.

Reference	Classifier	Corpus	Bully rate	F_1 score
[44]	SVM	1,762 tweets	39%	77%
[43]	wvec+SVM	1,762 tweets	39%	78%
[45]	smSDA+SVM	7,321 tweets	29%	72%
[45]	smSDA+SVM	1,539 MySpace posts	26%	78%

https://doi.org/10.1371/journal.pone.0203794.t010

These studies obtain higher scores on similar task but vastly different datasets. Notably, [45] shows a great improvement in classification performance using deep representational learning with a semantic-enhanced marginalized denoising auto-encoder over traditional *n*-gram and topic modelling features.

Conclusions and future research

The goal of the current research was to investigate the automatic detection of cyberbullying-related posts on social media. Given the information overload on the web, manual monitoring for cyberbullying has become unfeasible. Automatic detection of signals of cyberbullying would enhance moderation and allow to respond quickly when necessary.

Cyberbullying research has often focused on detecting cyberbullying 'attacks' and hence overlook other or more implicit forms of cyberbullying and posts written by victims and bystanders. However, these posts could just as well indicate that cyberbullying is going on. The main contribution of this paper is that it presents a system to automatically detect **signals of cyberbullying** on social media, including different types of cyberbullying, covering posts from bullies, victims and bystanders. We evaluated our system on a manually annotated cyberbullying corpus for English and Dutch and hereby demonstrated that our approach can easily be applied to different languages, provided that annotated data for these languages are available.

A set of binary classification experiments were conducted to explore the feasibility of automatic cyberbullying detection on social media. In addition, we sought to determine which information sources contribute most to the task. Two classifiers were trained on an English and Dutch ASKfm corpus and evaluated on a hold-out test of the same genre. Our experiments reveal that the current approach is a promising strategy for detecting signals of cyberbullying on social media automatically. After feature and hyperparameter optimisation of our models, a maximum F_1 score of 64.32% and 58.72% was obtained for English and Dutch, respectively. The classifiers hereby significantly outperformed a keyword and an (unoptimised) *n*-gram baseline. A qualitative analysis of the results revealed that false positives often include implicit cyberbullying or offenses through irony, the challenge of which will constitute an important area for future work. Error rates on the different author roles in our corpus revealed that especially victims are hard to recognise, as they react differently in our corpus, showing either powerlessness facing the bully or reacting in an assertive and sometimes even aggressive way.

As shown in [45] deep representation learning is a promising avenue for this task. We therefore intent to apply deep learning techniques to improve classifier performance.

Another interesting direction for future work would be the detection of fine-grained cyberbullying categories such as threats, curses and expressions of racism and hate. When applied in a cascaded model, the system could find severe cases of cyberbullying with high precision. This would be particularly interesting for monitoring purposes. Additionally, our dataset allows for detection of participant roles typically involved in cyberbullying. When applied as moderation support on online platforms, such a system enables feedback in function of the recipient (i.e. a bully, victim, or bystander).

Author Contributions

Conceptualization: Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, Véronique Hoste.

Data curation: Cynthia Van Hee, Gilles Jacobs, Els Lefever, Ben Verhoeven, Guy De Pauw.

Funding acquisition: Walter Daelemans, Véronique Hoste.

Methodology: Cynthia Van Hee, Gilles Jacobs, Bart Desmet, Els Lefever, Walter Daelemans, Véronique Hoste.

Project administration: Walter Daelemans, Véronique Hoste.

Resources: Cynthia Van Hee, Gilles Jacobs, Els Lefever.

Software: Cynthia Van Hee, Gilles Jacobs, Bart Desmet.

Supervision: Walter Daelemans, Véronique Hoste.

Writing – original draft: Cynthia Van Hee, Gilles Jacobs.

Writing – review & editing: Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, Véronique Hoste.

https://doi.org/10.1371/journal.pone.0203794

References

1. Livingstone S, Haddon L, Görzig A, Ólafsson K. Risks and safety on the internet: The perspective of European children. Initial Findings. London: EU Kids Online; 2010.
2. Tokunaga RS. Following You Home from School: A Critical Review and Synthesis of Research on Cyberbullying Victimization. Computers in Human Behavior. 2010;26(3):277–287. https://doi.org/10.1016/j.chb.2009.11.014
3. Mckenna KY, Bargh JA. Plan 9 From Cyberspace: The Implications of the Internet for Personality and Social Psychology. Personality & Social Psychology Review. 1999;4(1):57–75. https://doi.org/10.1207/S15327957PSPR0401_6
4. Gross EF, Juvonen J, Gable SL. Internet Use and Well-Being in Adolescence. Journal of Social Issues. 2002;58(1):75–90. https://doi.org/10.1111/1540-4560.00249
5. Juvonen J, Gross EF. Extending the school grounds?—Bullying experiences in cyberspace. Journal of School Health. 2008;78(9):496–505. https://doi.org/10.1111/j.1746-1561.2008.00335.x PMID:18786042
6. Hinduja S, Patchin JW. Bullies Move Beyond the Schoolyard: A Preliminary Look at Cyberbullying. Youth Violence And Juvenile Justice. 2006;4(2):148–169. https://doi.org/10.1177/1541204006286288
7. Van Cleemput K, Bastiaensens S, Vandebosch H, Poels K, Deboutte G, DeSmet A, et al. Zes jaar onderzoek naar cyberpesten in Vlaanderen, België en daarbuiten: een overzicht van de bevindingen. (Six years of research on cyberbullying in Flanders, Belgium and beyond: an overview of the findings.) (White Paper). University of Antwerp & Ghent University; 2013.
8. Livingstone S, Haddon L, Vincent J, Giovanna M, Ólafsson K. Net Children Go Mobile: The Uk report; 2014. Available from: http://netchildrengomobile.eu/reports. [Accessed 30th March 2018].
9. O'Moore M, Kirkham C. Self-esteem and its relationship to bullying behaviour. Aggressive Behavior. 2001;27(4):269–283. https://doi.org/10.1002/ab.1010
10. Fekkes M, Pijpers FIM, Fredriks AM, Vogels T, Verloove-Vanhorick SP. Do Bullied Children Get Ill, or Do Ill Children Get Bullied? A Prospective Cohort Study on the Relationship Between Bullying and Health-Related Symptoms. Pediatrics. 2006;117(5):1568–1574. https://doi.org/10.1542/peds.2005-0187 PMID:16651310
11. Cowie H. Cyberbullying and its impact on young people's emotional health and well-being. The Psychiatrist. 2013;37(5):167–170. https://doi.org/10.1192/pb.bp.112.040840
12. Price M, Dalgleish J. Cyberbullying: Experiences, Impacts and Coping Strategies as Described by Australian Young People. Youth Studies Australia. 2010;29(2):51–59.
13. Van Royen K, Poels K, Daelemans W, Vandebosch H. Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability. Telematics and Informatics. 2014;.
14. Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. 1995;20(3):273–297. https://doi.org/10.1007/BF00994018
15. Chang CC, Lin CJ. LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011;2(3):27:1–27:27.

16. Dalla Pozza V, Di Pietro A, Morel S, Emma P. Cyberbullying among Young People. European Parliament; 2016.

17. Olweus D. Bullying at School: What We Know and What We Can Do. 2nd ed. Wiley; 1993.

18. Smith PK, Mahdavi J, Carvalho M, Fisher S, Russell S, Tippett N. Cyberbullying: its nature and impact in secondary school pupils. Journal of Child Psychology and Psychiatry. 2008;49(4):376–385. https://doi.org/10.1111/j.1469-7610.2007.01846.x PMID:18363945

19. Vandebosch H, Van Cleemput K. Defining cyberbullying: a qualitative research into the perceptions of youngsters. Cyberpsychology and behavior: the impact of the Internet, multimedia and virtual reality on behavior and society. 2008;11(4):499–503. https://doi.org/10.1089/cpb.2007.0042

20. Vandebosch H, Van Cleemput K. Cyberbullying among youngsters: profiles of bullies and victims. New Media & Society. 2009;11(8):1349–1371. https://doi.org/10.1177/1461444809341263

21. Dooley JJ, Cross D. Cyberbullying versus face-to-face bullying: A review of the similarities and differences. Journal of Psychology. 2009;217:182–188.

22. Slonje R, Smith PK, Frisén A. The Nature of Cyberbullying, and Strategies for Prevention. Compututers in Human Behavior. 2013;29(1):26–32. https://doi.org/10.1016/j.chb.2012.05.024

23. Van Hee C, Verhoeven B, Lefever E, De Pauw G, Daelemans W, Hoste V. Guidelines for the Fine-Grained Analysis of Cyberbullying, version 1.0. LT3, Language and Translation Technology Team–Ghent University; 2015. LT3 15-01.

24. Salmivalli C, Lagerspetz K, Björkqvist K, Österman K, Kaukiainen A. Bullying as a group process: Participant roles and their relations to social status within the group. Aggressive Behavior. 1996;22(1):1–15. https://doi.org/10.1002/(SICI)1098-2337(1996)22:1%3C1::AID-AB1%3E3.0.CO;2-T

25. Vandebosch H, Van Cleemput K, Mortelmans D, Walrave M. Cyberpesten bij jongeren in Vlaanderen: Een studie in opdracht van het viWTA (Cyberbullying among youngsters in Flanders: a study commissoned by the viWTA). Brussels: viWTA; 2006. Available from: https://wise.vub.ac.be/fattac/mios/Eindrapport%20cyberpesten%20viwta%202006.pdf. [Accessed 30th March 2018].

26. Dehue F, Bolman C, Völlink T. Cyberbullying: Youngsters' Experiences and Parental Perception. Cyberpsychology and Behavior. 2008;11(2):217–223. https://doi.org/10.1089/cpb.2007.0008 PMID:18422417

27. Salmivalli C, Pöoyhönen V. In: Cyberbullying in Finland. 2nd ed. Wiley-Blackwell; 2012. p. 57–72.

28. Salawu S, he Y, Lumsden J. Approaches to Automated Detection of Cyberbullying: A Survey. IEEE Transactions on Affective Computing. forthcoming;p. 1.

29. Reynolds K, Kontostathis A, Edwards L. Using Machine Learning to Detect Cyberbullying. In: Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops. ICMLA'11. Washington, DC, USA: IEEE Computer Society; 2011. p. 241–244.

30. Dinakar K, Reichart R, Lieberman H. Modeling the Detection of Textual Cyberbullying. In: The Social Mobile Web. vol. WS-11-02 of AAAI Workshops. AAAI; 2011. p. 11–17.

31. Yin D, Davison BD, Xue Z, Hong L, Kontostathis A, Edwards L. Detection of Harassment on Web 2.0. In: Proceedings of the Content Analysis in the Web 2.0 (CAW2.0). Madrid, Spain; 2009.

32. Dadvar M. Experts and machines united against cyberbullying [PhD thesis]. University of Twente; 2014.

33. Nahar V, Al-Maskari S, Li X, Pang C. Semi-supervised Learning for Cyberbullying Detection in Social Networks. In: ADC.Databases Theory and Applications. Springer International Publishing; 2014. p. 160–171.

34. Bayzick J, Kontostathis A, Edwards L. Detecting the Presence of Cyberbullying Using Computer Software. In: Proceedings of the 3rd International Web Science Conference. WebSci11; 2011.

35. Sui J. Understanding and Fighting Bullying with Machine Learning [PhD thesis]. Department of Computer Sciences, University of Wisconsin-Madison; 2015.

36. Galán-García P, Puerta JGdl, Gómez CL, Santos I, Bringas PG. Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying. Logic Journal of the IGPL. 2016;24(1):42–53. Available from: http://dx.doi.org/10.1093/jigpal/jzv048.

37. Raisi E, Huang B. Cyberbullying Detection with Weakly Supervised Machine Learning. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. ASONAM'17. New York, NY, USA: ACM; 2017. p. 409–416.

38. Hosseinmardi H. Survey of Computational Methods in Cyberbullying Research. In: Proceedings of the First International Workshop on Computational Methods for CyberSafety. CyberSafety'16. New York, NY, USA: ACM; 2016. p. 4–4.

39. Nandhini B Sri, Sheeba JI. Online Social Network Bullying Detection Using Intelligence Techniques. Procedia Computer Science. 2015;45:485–492. https://doi.org/10.1016/j.procs.2015.03.085

40. Chavan VS, S SS. Machine learning approach for detection of cyber-aggressive comments by peers on social media network. In: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI); 2015. p. 2354–2358.

41. Van Hee C, Lefever E, Verhoeven B, Mennes J, Desmet B, De Pauw G, et al. Detection and fine-grained classification of cyberbullying events. In: Angelova G, Bontcheva K, Mitkov R, editors. Proceedings of Recent Advances in Natural Language Processing, Proceedings; 2015. p. 672–680.

42. Van Royen K, Poels K, Vandebosch H. Harmonizing freedom and protection: Adolescents' voices on automatic monitoring of social networking sites. Children and Youth Services Review. 2016;64:35–41. https://doi.org/10.1016/j.childyouth.2016.02.024

43. Zhao R, Zhou A, Mao K. Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features. In: Proceedings of the 17th International Conference on Distributed Computing and Networking. No. 43 in ICDCN'16. New York, NY, USA: ACM; 2016. p. 43:1–43:6.

44. Xu JM, Jun KS, Zhu X, Bellmore A. Learning from Bullying Traces in Social Media. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL HLT'12. Stroudsburg, PA, USA: Association for Computational Linguistics; 2012. p. 656–666.

45. Zhao R, Mao K. Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder. EEE Transactions on Affective Computing. 2017;8(3):328–339. https://doi.org/10.1109/TAFFC.2016.2531682

46. Huang Q, Singh VK, Atrey PK. Cyber Bullying Detection Using Social and Textual Analysis. In: Proceedings of the 3rd International Workshop on Socially-Aware Multimedia. SAM'14. New York, NY, USA: ACM; 2014. p. 3–6.

47. Al-garadi MA, Dewi Varathan K, Devi Ravana S. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. Computers in Human Behavior. 2016;63:433–443. https://doi.org/10.1016/j.chb.2016.05.051

48. Squicciarini A, Rajtmajer S, Liu Yh, Griffin C. Identification and Characterization of Cyberbullying Dynamics in an Online Social Network. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. ASONAM'15. New York, NY, USA: ACM; 2015. p. 280–285.

49. Chatzakou D, Kourtellis N, Blackburn J, De Cristofaro E, Stringhini G, Vakali A. Mean Birds: Detecting Aggression and Bullying on Twitter. In: Proceedings of the 2017 ACM on Web Science Conference. WebSci'17. New York, NY, USA: ACM; 2017. p. 13–22.

50. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer PW. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research (JAIR). 2002;16:321–357. https://doi.org/10.1613/jair.953

51. Van Hee C, Lefever E, Verhoeven B, Mennes J, Desmet B, De Pauw G, et al. Automatic detection and prevention of cyberbullying. In: Lorenz P, Bourret C, editors. International Conference on Human and Social Analytics, Proceedings. IARIA; 2015. p. 13–18.

52. Hinduja S, Patchin JW. Cyberbullying: Neither an epidemic nor a rarity. European Journal of Developmental Psychology. 2012;9(5):539–543. https://doi.org/10.1080/17405629.2012.706448

53. Slonje R, Smith PK. Cyberbullying: Another main type of bullying? Scandinavian Journal of Psychology. 2008;49(2):147–154. https://doi.org/10.1111/j.1467-9450.2007.00611.x PMID:18352984

54. O'Sullivan PB, Flanagin AJ. Reconceptualizing 'flaming' and other problematic messages. New Media & Society. 2003;5(1):69–94. https://doi.org/10.1177/1461444803005001908

55. Willard NE. Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. 2nd ed. Research Publishers LLC; 2007.

56. Bastiaensens S, Vandebosch H, Poels K, Van Cleemput K, DeSmet A, De Bourdeaudhuij I. Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. Computers in Human Behavior. 2014;31:259–271. https://doi.org/10.1016/j.chb.2013.10.036

57. Salmivalli C. Bullying and the peer group: A review. Aggression and Violent Behavior. 2010;15(2):112–120. https://doi.org/10.1016/j.avb.2009.08.007

58. Bastiaensens S, Vandebosch H, Poels K, Van Cleemput K, DeSmet A, De Bourdeaudhuij I. 'Can I afford to help?' How affordances of communication modalities guide bystanders' helping intentions towards harassment on social network sites. Behaviour & Information Technology. 2015;34(4):425–435. https://doi.org/10.1080/0144929X.2014.983979

59. Salmivalli C, Voeten M, Poskiparta E. Bystanders Matter: Associations Between Reinforcing, Defending, and the Frequency of Bullying Behavior in Classrooms. Journal of Clinical Child & Adolescent Psychology. 2011;40(5):668–676. https://doi.org/10.1080/15374416.2011.597090

60. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. brat: a Web-based Tool for NLP-Assisted Text Annotation. In: Proceedings of the Demonstrations Session at EACL 2012. Avignon, France; 2012. p. 102–107.

61. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960;20(1):37–46. https://doi.org/10.1177/001316446002000104

62. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971;76(5):378–382. https://doi.org/10.1037/h0031619

63. McHugh ML. Interrater reliability: the kappa statistic. Biochemia Medica. 2012;22(3):276–282. https://doi.org/10.11613/BM.2012.031 PMID:23092060

64. He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering. 2009;21(9):1263–1284. https://doi.org/10.1109/TKDE.2008.239

65. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research. 2008;9:1871–1874.

66. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

67. Desmet B. Finding the online cry for help: automatic text classification for suicide prevention [PhD thesis]. Ghent University; 2014.

68. Hoste V. Optimization Issues in Machine Learning of Coreference Resolution [PhD thesis]. Antwerp University; 2005.

69. van de Kauter M, Coorman G, Lefever E, Desmet B, Macken L, Hoste V. LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. Computational Linguistics in the Netherlands Journal. 2013;3:103–120.

70. Osgood CE, Walker EG. Motivation and language behavior: A content analysis of suicide notes. The Journal of Abnormal and Social Psychology. 1959;59(1):58. https://doi.org/10.1037/h0047078

71. Jijkoun V, Hofmann K. Generating a Non-English Subjectivity Lexicon: Relations That Matter. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA; 2009. p. 398–405.

72. De Smedt T, Daelemans W. "Vreselijk mooi!" ("Terribly Beautiful!"): A Subjectivity Lexicon for Dutch Adjectives. In: Proceedings of the Eight International Conference on Language Resources and Evaluation. LREC'12. Istanbul, Turkey; 2012. p. 3568–3572.

73. Hu M, Liu B. Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining. KDD04. ACM; 2004. p. 168–177.

74. Wilson T, Wiebe J, Hoffmann P. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. HLT'05. Association for Computational Linguistics; 2005. p. 347–354.

75. Stone PJ, Dunphy DCD, Smith MS, Ogilvie DM. The General Inquirer: A Computer Approach to Content Analysis. The MIT Press; 1966.

76. Nielsen FÅ. A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. In: Rowe M, Stankovic M, Dadzie AS, Hardey M, editors. Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages. vol. 718 of CEUR Workshop Proceedings. CEUR-WS.org; 2011. p. 93–98.

77. Mohammad S, Dunne C, Dorr B. Generating High-coverage Semantic Orientation Lexicons from Overtly Marked Words and a Thesaurus. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2. EMNLP'09. Stroudsburg, PA, USA: Association for Computational Linguistics; 2009. p. 599–608.

78. Pennebaker JW, Francis ME, Booth RJ. Linguistic Inquiry and Word Count: LIWC 2001. Mahwah, NJ: Lawrence Erlbaum Associates; 2001.

79. Zijlstra H, Van Meerveld T, Van Middendorp H, Pennebaker JW, Geenen R. De Nederlandse versie van de 'linguistic inquiry and word count' (LIWC). Gedrag Gezond. 2004;32:271–281.

80. Rehurek R, Sojka P. Software framework for topic modelling with large corpora. In: The LREC 2010 Workshop on new Challenges for NLP Frameworks. University of Malta; 2010. p. 45–50.

81. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of Machine Learning Research. 2003;3:993–1022.

82. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of the American Society for Information Science. 1990;41:391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9

83. Baroni M, Bernardini S. BootCaT: Bootstrapping Corpora and Terms from the Web. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation. LREC'04; 2004. p. 1313–1316.

84. Fawcett T. An introduction to ROC analysis. Pattern recognition letters. 2006;27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010

5.7 Case Study 2

RESEARCH ARTICLE

Ten quick tips for staying safe online

Danielle Smalls[1☯], **Greg Wilson**[2☯]

1 MIR Community Group, Raleigh, North Carolina, United States of America, 2 RStudio PBC, Boston, Massachusetts, United States of America

☯ *These authors contributed equally to this work.*

Introduction

Researchers studying everything from sexual health to the Coronavirus Disease 2019 (COVID-19) to gun violence are increasingly likely to be targeted because of their work. While research institutions have rules and guidelines for safeguarding sensitive information, these usually do not address the problem of keeping individuals safe from either targeted attacks like Climategate [1] or the kinds of "drive-by" threats that everyone now faces regardless of their occupation.

Hollywood depictions of everyday threats are as far from reality as their portrayals of scientists, but more realistic guidance for personal digital security is now freely available [2–4]. The 10 quick tips in this paper are a starting point: While they apply to everyone, they were developed with researchers in mind. While researchers expect their work to be scrutinized by the academic community, they should not expect to endure harassment due to the visibility of their published works. These rules do not guarantee complete safety, any more than seatbelts guarantee safe driving, but following them greatly reduces the likelihood of harm.

Rule 1: Put on your own mask

The first and most important rule is that we should not rely on companies, universities, and other institutions to protect us, for the simple reason that they are not penalized if they don't. As recently as 10 years ago, we could blame the lack of meaningful institutional liability for data breaches on the law being slow to catch up with rapidly changing technology. Accountability for these breaches is practically nonexistent: Data breaches have minimal impact on companies' profitability, and individuals are almost never fined, much less jailed.

Much of what institutions force us to go through is **security theater** intended to make us believe something is being done rather than to actually make us safer. Requiring people to take off their shoes at airports is 1 example; random searches of backpacks and bags at the entrance to the subway is another, since it's hard to imagine that a would-be attacker wouldn't just go to another entrance. (Bruce Schneier's blog [5] has many examples of security theater and the harm it does.)

Security theater is counterproductive because it encourages us to cut corners in ways that actually make us less safe. For example, forcing people to change passwords every 3 months encourages people to choose memorable (and therefore easy-to-guess) passwords.

Rule 2: Digital security is rarely the weakest link

The second rule is to remember that most attacks take place offline and that the most effective ones are often the simplest. At an airport several years ago, 1 author heard a professor of computer science try to reset an

https://doi.org/10.1371/journal.pone.1008563 Used with CC 4.0 license

online account over the phone. In just a couple of minutes, they had inadvertently told everyone in the lounge their full name, their date of birth, the 3-digit verification code on the back of their credit card, and what was almost certainly their mother's maiden name.

The moral of this story is that safety comes from good habits, not technology. **Social engineering** is far more common than hacking: In practice, it is far easier to trick someone into giving you their password than to break into their devices digitally.

The key practice is **situational awareness**, which is a fancy way of saying, "Pay attention to what's happening and respond accordingly." If you start working on a high-profile subject that will likely attract controversy, you should take more precautions than usual. For example, someone should recognize that agreeing to be an expert witness increases the odds that they will be targeted and should be more careful about what he puts into email while preparing and delivering his testimony.

The corollary to situational awareness is to de-escalate when you can. Being on guard all the time is exhausting and quickly leads to **security fatigue** [6]. If you are too tired to follow good practices, knowing them does you no good.

Two technologies that are useful, but only to a certain extent, are virtual private networks (VPNs) and a specialized web browser called "Tor." A VPN connects your device to a server, then has the server make connections to other machines on your behalf. All messages between your device and the server are encrypted, and the server can be managed by professional IT staff in a jurisdiction with tight privacy laws to increase your safety. Tor routes messages randomly through a network of servers, making traffic much harder to track. Both reduce risk, but neither eliminates it if your device is compromised, if the VPN is subpoenaed, or if you log in to accounts over Tor (thereby revealing your identity to those sites).

Rule 3: Use relevant threat models

Edward Snowden and the journalists who worked with him took extraordinary measures to safeguard themselves against **state-level actors** [7], but most of us aren't involved in issues of national security and don't need to take those kinds of precautions. Instead, we typically face 1 of 3 kinds of threat illustrated by the examples below.

- **Casual threats** are opportunistic. For example, Monica, a professor in psychology, is targeted by Mohan, an undergraduate in computer science who spends hours every day in online echo chambers complaining about how "SJW bullshit" is ruining tech. He really didn't enjoy Monica's guest lecture on discrimination and inclusivity in his software engineering class and thinks it would be a laugh to make her the target of anonymous abuse online. He is unlikely to invest significant effort in his attack (at least not initially), but his attack may be backed up by more knowledgeable members of online forums. They are almost certainly not computer security specialists; instead, they are probably older versions of Mohan who have picked up a few tricks and bits of software and enjoy the digital equivalent of throwing bricks through strangers' windows.
- **Intimate threats** come from people who know their targets' passwords or have a chance to install spyware on their targets' devices [8]. For example, Elena, graduate student, is targeted by her former romantic partner Eric, who is also a graduate student but not in the same department. Their relationship had become increasingly abusive over the last 2 years. With the help of friends, Elena has moved out of their shared apartment and is rebuilding her life; Eric is obsessed with the idea that she left him for someone else and is now stalking her.

https://doi.org/10.1371/journal.pone.1008563

- **Insider threats** come from people who have legitimate access to accounts and devices. For example, Boris, professor of medicine, is targeted by Bethany, who works for the university's IT department. Boris has agreed to serve as an expert witness in an upcoming liability case involving a large chemical company; Bethany has been asked by a former colleague to find out what he is going to say in order to discredit his testimony.

Rule 4: Use a password manager

Using a weak password is a good way to ensure that your account will eventually be compromised, in part because **dictionary attacks** can be run offline against encrypted password files to find passwords that match common patterns. Using a clever password scheme, such as the name of the site plus a word only you know, does not increase security by much: Whatever scheme you have thought of, attackers have seen before. And since people are often identified on multiple sites by the same email address, as soon as 1 site where you've used that scheme is compromised, attackers can guess the scheme and use it elsewhere.

Reusing passwords ensures that damage spreads, so using a different password for each site helps limit harm if any are compromised. However, strong passwords are hard to remember and to type, so always use a **password manager** that generates strong passwords and saves them all under a master **passphrase**. Your passphrase should be several words long and something you are unlikely to forget. It does create a single point of attack, but is still safer than choosing passwords yourself, since password managers aren't fooled by similar-seeming sites like paypaI.com.

> Writing passwords down and keeping them in your wallet is not necessarily a bad practice—it depends on who is doing it. For example, an elderly person who finds tech confusing might well choose simple, easy-to-guess passwords for their accounts if they have to be remembered. On the other hand, they have a lifetime of practice keeping track of bits of paper, and will probably notice if their purse or wallet is stolen.

Rule 5: Use two-factor authentication

Authentication relies on something you know (like a password), something you have (like a security key), or something you are (like your fingerprints). **Two-factor authentication** (2FA) requires two of these together to establish your identity, e.g., a password (which can be stolen electronically) plus a random code generated by an app on your phone (which means attackers need access to you).

2FA is as important to security as using a password manager, but where possible, you should rely on an app for 2FA instead of using text messages. What you should never do is share a confirmation code, since a common attack is to trigger a password reset and then call the victim pretending to be from the IT department and ask them to read the code back to "verify" your account. As soon as you do this, the attacker can change your password and get into your account.

> Many security experts now recommend using a physical 2FA key such as a YubiKey, which fits on a keychain and plugs into a standard USB port. Sites like Tech Solidarity have easy-to-follow instructions explaining how to set them up for a range of popular social networking sites.

Rule 6: Think before opening

Much of the software we use was designed in more innocent times, and since companies are almost never held liable for the damage caused by their software, they have consistently prioritized convenience for the many over harm to the few. One common example is documents that contain code called "macros" that

https://doi.org/10.1371/journal.pone.1008563

automatically execute when the document is opened. Used for good, a macro can check that an address field has been filled in correctly. Used for evil, it can email everyone in your address book or send a copy of those addresses to anyone in the world. Microsoft Word and Excel are particularly notorious for this vulnerability, but many other kinds of documents have the same flaw.

Attempts to get you to open an email attachment, click on a link, install software, or log into a website are called **phishing** attacks. The strongest defense is to never do these things, but in the modern world, that would make most work impossible. The second-best defense is to take sensible precautions. If you are able, invest in virus scanning software such as Proofpoint to scan email attachments before you download them. While many email clients have virus scanning technology built-in, this will offer an extra layer of protection.

Similarly, don't click links in emails without checking them first: Instead, hover over the link and see if it matches the site it claims to be. Alternatively, log into the site manually rather than following the provided link. It takes more time, but is still faster than fixing your credit rating. And when you go to a website, check the real domain name in the URL: paypaI.com with an upper-case "I" instead of a lower-case "l" is not the site it pretends to be, and wwwpaypal.com is a different domain than www.paypal.com.

> Many sites send an email with a random URL to confirm your identity when you are resetting your password. On the one hand, this means that an attacker has to get access to your email in order to break into your account. On the other hand, random URLs are hard to type in, so these emails encourage us to click on links in emails. If you are not expecting a password reset email, don't click on the link.

While phishing attacks are wide-ranging, **spear phishing** describes the use of data harvested from previous victims to attack specific targets. Here, the best defense is to verify suspicious emails, e.g., by phoning people to confirm their identity. It's particularly important to do this when you are sent things like password reset instructions. Many IT departments send out messages that are indistinguishable from spear phishing attacks, which just trains people to be victims.

Rule 7: Erase before discarding

Moving files into the trash and then emptying it does not actually erase the data: It just tells the computer that the space is available for reuse. (This is why reporters and private investigators regularly go dumpster diving.) The best way to address this problem is to encrypt your hard drive, which is a quick setup option for all major operating systems these days.

Even with that, you should act as if any device you throw away is going to fall into unfriendly hands. Use a secure deletion tool like BleachBit (Linux or Windows) or FileShredder (MacOS) before selling, recycling, or discarding your hardware, but keep in mind that this doesn't affect backups or files stored online on sites like Dropbox. And keep in mind that it is practically impossible to truly delete data from social networking sites: In most cases, their "delete" usually means "don't show any more" rather than "erase all past record of."

Rule 8: Check your devices and accounts periodically

Many tech companies who offer free products and services make money by selling targeted advertising to you using the data they have about you. They do give users some control over personal data, but they frequently change their terms of service in opaque ways. Seemingly innocuous information can give attackers valuable clues: Restaurant "likes" reveal where you were at specific times, while funny stories about childhood birthday parties reveal likely answers to security questions. Again, it's a good practice to get into the habit of checking your privacy settings every time you do some other regular task.

https://doi.org/10.1371/journal.pone.1008563

Unfortunately, even if you do this, information may leak through other means. For example, attackers can friend your friends in an attempt to get information about you, such as the name of your first school. And as bad as social media sites are for social engineering in this way, cell phone applications are often worse (not even counting the ones that turn out to be government-sponsored spyware [9]). In general, if a game wants access to your camera and address book, you should probably find a different game to play.

Since social media is a fact of life for most of us, you should check your settings periodically, just as you would take your car in for an oil change. (The authors do these things at the same time in order to remember both.) Turn off everything you can, and then use a tracking blocker such as Ghostery to reduce information leakage.

Many experts recommend using separate devices or accounts for work and personal life, but this is increasingly unrealistic. Everyone checks their personal email from their work device eventually, and everyone uses their personal phone for 2FA. However, you should consider getting a second phone for international travel: The legalities around who can take your devices and/or force you to unlock them are complicated and frequently overstepped, so you should assume that anything on or connected to the devices you are traveling with will be compromised.

> Never plug a random USB drive into your device: it's like letting a complete stranger into your home unsupervised.

Rule 9: Fight back

Casual attackers may eventually get bored and move on, but like all bullies, they will also often revisit previous victims, and even if they don't, they are likely to pick new ones. If you have been attacked:

1. **Find support.** Being targeted is frightening and wearying, particularly if you belong to one of the many groups that are targeted in real life as well as online. Let family, friends, and colleagues know what is happening so that they can support you. They may also be able to offer advice if they have been in similar situations.
2. **Use anti-harassment apps** like Block Party and document everything. Save emails and take screenshots of sites like Facebook and Twitter (in case attackers delete or alter material).
3. **Do not engage directly.** Casual attackers are often seeking attention, so a direct response often encourages further attacks (and can draw attention from like-minded attackers).
4. **Report the attack.** Social media sites have done everything they can to avoid legal accountability for online attacks, but companies and universities will usually take what steps they can once they know there is a problem. In the authors' experience, they are more inclined to take real action against the attacker if they believe that you might speak publicly about what has happened and thereby damage their reputation, so never agree to a nondisclosure agreement that would prevent you from doing so.

Rule 10: It's not all about you

Our final rule brings us full circle to the first one. We don't just wear masks to prevent ourselves from becoming infected: We also wear them so that we will not infect others. Similarly, if you do not take precautions with online security, then you are putting others at risk. Simple steps like putting passwords on PDFs that contain sensitive information can go a long way to deter attackers, in the same way that a sturdy-looking bike lock encourages would-be thieves to go after some other bike. And if you are compromised, let those affected know as soon as you can.

https://doi.org/10.1371/journal.pone.1008563

The only long-term way to improve everyone's online safety is to pressure politicians to strengthen liability legislation so that companies, universities, and other institutions have real incentives to take meaningful action. Cars and drugs are as safe as they are because their manufacturers are liable for negligence and harm. The sooner software companies and social media sites are liable as well, the safer all of us will be.

Acknowledgments

The authors are grateful to Claire Bowen (Urban Institute), Leigh Honeywell (Tall Poppy), Mike Hoye (Mozilla), Scott Jackson (RStudio), Rick Johnson (RStudio), Cheng Soon Ong (CSIRO), and Sydney Young (EFF) for their feedback on this article.

References

1. The Editors. Closing the Climategate. Nature. 2010;468(7322):345–5. https://doi.org/10.1038/468345a PMID:21085128
2. Front Line Defenders. Digital Security Resources. [cited 2020 Jan 25]. Available from: https://www.frontlinedefenders.org/en/digital-security-resources.
3. European Federation of Journalists. Cyber security training for journalists: Tips and tricks to keep your communications safe. [cited 2020 Jan 25]. Available from: https://europeanjournalists.org/blog/2015/01/22/cyber-security-training-for-journalists/.
4. Electronic Frontier Foundation. Surveillance Self-Defense: Tips, Tools and How-Tos for Safer Online Communications. [cited 2020 Jan 25]. Available from: https://ssd.eff.org/.
5. Schneier B. Schneier on Security. [cited 2020 Nov 3]. Available from: https://www.schneier.com/.
6. Stanton B, Theofanos MF, Prettyman SS, Furman S. Security Fatigue. IT Prof. 2016;18(5):26–32. https://doi.org/10.1109/mitp.2016.84
7. Snowden E. Permanent Record. Metropolitan Books; 2019.
8. Leitão R. Technology-Facilitated Intimate Partner Abuse: a qualitative analysis of data from online domestic abuse forums. Hum Comput Interact. 2019:1–40. https://doi.org/10.1080/07370024.2019.1685883
9. Schneier B. ToTok Is an Emirati Spying Tool. [cited 2020 Jan 25]. Available from: https://www.schneier.com/blog/archives/2019/12/totok-is-an-emi.html.

5.8 Case Study 3

RESEARCH ARTICLE

Warning: Humans cannot reliably detect speech deepfakes

Kimberly T. Mai[1,2]*, Sergi Bray[1,2], Toby Davies[1], Lewis D. Griffin iD[2]

1 Department of Security and Crime Science, University College London, London, United Kingdom,

2 Department of Computer Science, University College London, London, United Kingdom

* *kimberly.mai@ucl.ac.uk*

Abstract

Speech deepfakes are artificial voices generated by machine learning models. Previous literature has highlighted deepfakes as one of the biggest security threats arising from progress in artificial intelligence due to their potential for misuse. However, studies investigating human detection capabilities are limited. We presented genuine and deepfake audio to n = 529 individuals and asked them to identify the deepfakes. We ran our experiments in English and Mandarin to understand if language affects detection performance and decision-making rationale. We found that detection capability is unreliable. Listeners only correctly spotted the deepfakes 73% of the time, and there was no difference in detectability between the two languages. Increasing listener awareness by providing examples of speech deepfakes only improves results slightly. As speech synthesis algorithms improve and become more realistic, we can expect the detection task to become harder. The difficulty of detecting speech deepfakes confirms their potential for misuse and signals that defenses against this threat are needed.

Introduction

Adversaries are already using speech deepfakes to commit fraud. In 2020, a bank manager in Hong Kong received a phone call from someone sounding like a company director he had spoken to before [1]. The purported director requested the bank manager to authorize transfers totaling $35 million. Based on their existing relationship, the bank manager transferred $400,000 until he realized something was wrong. The bank manager was a victim of an elaborate hoax: fraudsters had used deepfake technology to clone the director's voice. This incident is not isolated. In 2019, the CEO of a UK-based firm was swindled by a speech deepfake of his manager into transferring €220,000 to a Hungarian supplier [2].

Speech deepfakes are artificial voices generated by machine learning models. Due to rapid research progress, it is possible to produce a realistic-sounding clone using only a few audio samples [3]. This development raises the prospect of exploiting speech deepfakes for various criminal activities. Alongside impersonation, criminals may use deepfakes for spear phishing, propagating fake news, and bypassing biometric authentication systems [4–6].

Existing speech deepfake detection research focuses on developing machine learning systems in the context of voice authentication [7–9]. Comparisons beyond this biometric setting and studies which measure human detection capabilities are sparse [10].

The state of existing research raises questions. Firstly, machine learning systems require large amounts of data for training [11] and are hard to interpret [12]. When analyzing these systems, it is unclear which

https://doi.org/10.1371/journal.pone.0285333 Used with CC 4.0 license

characteristics distinguish synthesized speech from bona fide. Therefore, knowing what humans use to identify deepfakes could provide a better understanding of how black-box machine learning systems work.

Secondly, focusing on automated biometric authentication does not quantify the threat of other potential criminal applications of speech deepfakes. Understanding the extent of these threats is critical. Multiple studies deem other uses of speech deepfakes as more concerning, such as misleading people through voice impersonations [5, 6]. Experts expect disinformation from deepfakes to erode trust on several levels: towards individuals, organizations, and even societies [13]. Moreover, it is estimated as much as 90% of online content will be synthetically generated by 2026 [14], meaning it will be challenging to moderate what gets produced. Therefore, understanding the risks of speech deepfakes will enable the development of better defenses and regulations to counteract hazards before they occur.

We seek to address these two questions by measuring how well humans distinguish bona fide speech from synthesized speech. We ran an online experiment where individuals listened to bona fide and fake audio clips and attempted to differentiate between them.

We randomly assigned the participants to two configurations. In the first configuration, we presented participants with one audio clip at a time and asked them to decide if the clip was fake. In the second configuration, we presented participants with audio clip pairs containing the same speech (one bona fide and one synthesized) and asked them to identify the synthesized audio.

We ran the experiment in English and Mandarin to understand if listeners used language-specific attributes to detect deepfakes and to observe if deepfake detection is more manageable in one language than another. Finally, we incorporated randomized interventions to evaluate whether familiarizing participants with examples of speech deepfakes boosts detection performance.

Our results suggest the listeners had limited detection capabilities, and performance is similar between languages. Additionally, familiarizing participants improved performance but only to a small extent.

Background

Deepfake media

Deepfakes are synthetic media produced in the likeness of a person. They fall under the field of generative artificial intelligence (AI). Generative AI is a subset of machine learning (ML) algorithms that learn the patterns and characteristics of a dataset [11]. The algorithms use this knowledge to generate synthetic content similar to the original data. Deepfakes specifically refer to the outputs of generative AI that resemble humans and their actions.

Deepfake media occur in different modalities:

1. **Images**: This modality contains static faces generated using varying techniques. These techniques include:
    - **Generation from scratch**: A generative adversarial network [15] or diffusion model [16] synthesizes a fictional identity.
    - **Morphing**: Blending similar-looking faces to produce an identity containing the characteristics of the sources [17].
    - **Swaps**: A source face replaces the target in a different image [18].

https://doi.org/10.1371/journal.pone.0285333

2. **Video**: This modality features individuals performing actions. Currently, the techniques used to synthesize videos are similar to those used in images. Image synthesis techniques are applied at a frame level and stitched together to form a video.
3. **Speech**: This modality conveys information in a manner that sounds like a genuine person's voice. Although audio can refer to general sound synthesis, the terms "audio," "speech," and "voice" deepfakes are used interchangeably in academic literature. We refer to them as "speech deepfakes" for consistency throughout the text.

In addition, deepfakes are either produced in the likeness of a known identity (**targeted**) or do not resemble a familiar identity (**untargeted**). For example, we can categorize video deepfakes of politicians as targeted. Conversely, a generic face created from scratch and not conditioned to resemble a specific individual is untargeted.

We refer the reader to Zhang (2022) [19] for further information on deepfake terminology. As fewer works focus on speech deepfakes, we concentrate on this modality.

Synthesizing speech

Generative models are often used to synthesize speech. Speech synthesizers which use generative models follow a common framework:

1. **Data collection**: Several audio recordings of the speaker are collected.
2. **Pre-processing**: The audio recordings are converted into alternative formats to make it easier for the generative model to work with them.
3. **Training**: Processed audio recordings are fed to the generative model to learn the patterns and characteristics of the data. The trained model is often called a vocoder.

The frameworks often include text-to-speech (TTS) modules to make it easier to generate speech. The generative model also sees text transcriptions corresponding to the audio recordings in this setting.

We depict a visualization of this framework in Fig 5.1.

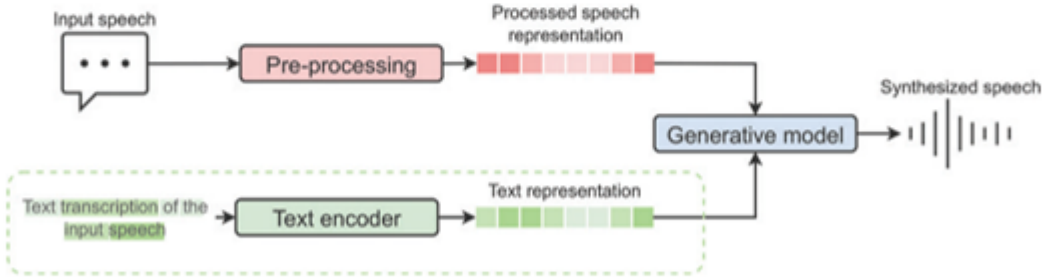


Fig 5.1 Diagram of a typical generative speech synthesis model.
https://doi.org/10.1371/journal.pone.0285333.g001

Related work on human deepfake detection capabilities

Most deepfake detection studies which examine human performance use visual media. When faced with deepfake content of politicians, participants rely on contextual knowledge in the form of political literacy to identify spoofs [20, 21].

Removing such background knowledge makes the detection task more difficult. In the context of images, multiple studies show humans do not perform much better than chance [22, 23]. There is no improvement when evaluating videos either [24–26]. Moreover, these studies suggest humans are overconfident in their deepfake detection abilities [25].

Several of the above studies examine if interventions can boost detection performance. However, the effectiveness of these interventions is debatable. Bray et al. [22] familiarized participants by showing examples of deepfakes before the main task. The authors also drew participants' attention to errors often present in bogus images. Although these interventions improved deepfake detection performance, they also increased overall skepticism as a higher proportion of bona fide images were falsely classified. One could also note that pointing out errors biases the participants and prevents them from independently identifying the tell-tale characteristics of deepfakes. Köbis et al. [25] presented interventions by informing participants about the impact of deepfakes and rewarding correct guesses. Neither intervention led to improved performance.

In contrast, other authors found interventions derived from ML model outputs improve detection. Tahir et al. [26] produced educational material containing indicators of bogus images with the assistance of ML interpretability tools. The authors found detection performance improved compared to the initial control group. However, a recent study [27] contests the reliability of these tools, as the authors show it is possible to manipulate the output visualizations. Groh et al. [24] allowed participants to amend their choices after viewing the predictions of an ML model. This form of cooperation improved results significantly.

There are fewer studies that examine how well humans can detect speech deepfakes. Watson et al. [28] presented eight clips to college students and asked them to decide whether the clips were real or fake. They found that shorter clips were easier to identify. However, the sample size of their study was small and skewed towards a younger, college-educated demographic. The ASVspoof challenge organizers ran an experiment with a larger sample size [29]. They asked 1,145 participants to imagine they worked in a call center and decide whether the incoming calls were spoken by humans or by an AI. However, the experiment is limited to the speaker verification setting.

Müller et al. [30] ran a game where 378 participants competed against an ML model to decide if an audio clip was fake. Similarly to Groh et al. [24], they found that feedback from the ML model improved human performance. In their experiment, Müller et al. [30] found that the difference between human and AI accuracy was about 10%. However, their study only used English-language clips, only presented one audio clip to participants at a time, and did not collect information about participant confidence.

We summarize the relevant literature in Table 5.11. We note that Barari et al. [20] mention fake speech stimuli in their analysis. However, they used actors to create the speech instead of generative AI. Therefore we excluded this from our analysis.

https://doi.org/10.1371/journal.pone.0285333

Table 5.11 Summary of related literature measuring human capabilities to detect deepfakes.

Modality	Year	Author	Deepfake stimuli
Image	2021	Nightingale & Farid [23]	Faces generated using StyleGAN2 [31]
	2022	Bray et al. [22]	Faces generated using StyleGAN2
Video	2021	Barari et al. [20]	Face-swap videos of politicians
	2021	Groh et al. [24]	Face-swap videos from the Deepfake Detection Challenge dataset [32]
	2021	Köbis et al. [25]	Face-swap videos from the Deepfake Detection Challenge dataset
	2021	Tahir et al. [26]	Face-swap videos from Celeb-DF [33], FaceForensics++ [34] and DeepFaceLab [35]
	2022	Appel & Prietzel [21]	Face-swap videos of politicians
Speech	2020	Wang et al. [29]	Spoofed utterances generated from TTS and voice conversion systems used in ASVspoof 2019
	2021	Watson et al. [28]	Audio clips generated using MelGAN [36]
	2022	Müller et al. [30]	Spoofed utterances from the ASVspoof 2019 dataset [29]

https://doi.org/10.1371/journal.pone.0285333.t001

Materials and methods

Our research questions were as follows:

1. How well can humans detect speech deepfakes?
2. Are there differences in detection capabilities depending on the language?
3. Do interventions in the form of examples and added context improve detection performance?

Our experiments focused on human performance rather than the performance of automated detectors. Through this setup, we could quantify the threat of speech deepfakes when humans interact with them.

Stimuli

Bona fide stimuli. We collected bona fide stimuli from two publicly available datasets. Both datasets consist of one female speaker reading generic sentences. The datasets also include text transcriptions of the audio. We chose such datasets to prevent participants from using external cues for the detection task.

We used LJSpeech [37] as the English dataset. The dataset consists of a speaker reading passages from seven non-fiction books, varying between one and ten seconds in length.

We used the Chinese Standard Mandarin Speech Corpus (CSMSC) [38] as the Mandarin dataset. The corpus used in the dataset aims to cover Mandarin tones and prosody as comprehensively as possible.

Deepfake stimuli. To create the deepfake stimuli, we used publicly available TTS models trained on the two datasets [39]. In particular, we chose pre-trained VITS models [40]. VITS is an end-to-end TTS model which combines the data pre-processing and vocoder into a single framework.

We randomly selected 50 sentences from the validation split of the two datasets to create the deepfakes. We used the same sentences for our bona fide stimuli. Therefore, we had 100 clips in total.

https://doi.org/10.1371/journal.pone.0285333

The validation split consists of samples not used for training the ML models. It is good practice to use unseen data because it indicates how well a trained model generalizes. Consequently, the resultant generated audio should contain artifacts that we would expect to hear from ML models. These artifacts might serve as informative features for distinguishing deepfakes. If we used samples previously seen during training, the model could potentially mimic the samples perfectly and would not contain representative artifacts.

Procedure

The setup for the English and Mandarin experiments was identical. We randomly assigned participants to two configurations: unary and binary. In both configurations, we asked participants to rate the confidence of their choice on a ten-point Likert scale and provide freeform text justifications. Participants were allowed to listen to the clips as often as they liked. We did not give feedback to the participants to inform them if their choices were correct. Compared to the setups described in Müller et al. and Groh et al. [24, 30], the lack of feedback creates a more realistic scenario. When encountering speech deepfakes in the wild (for example, through fraudulent calls), humans do not know that the voices are fake. We include screenshots of the two configurations in Fig 5.2.

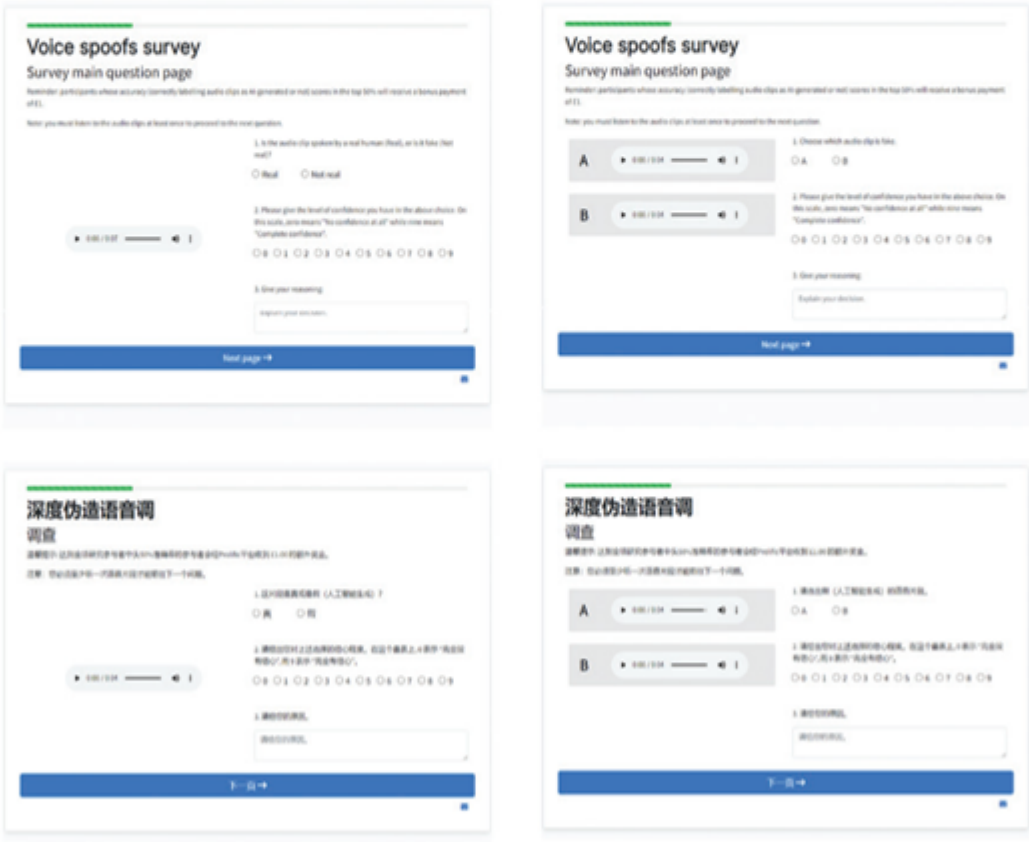


Fig 5.2 Screenshots of the task interface.

https://doi.org/10.1371/journal.pone.0285333.g002

Unary. We presented 20 randomly chosen distinct clips to each participant, each on separate pages. Participants listened to approximately an equal number of bona fide and synthesized clips, but we did not inform them about the proportion. We tasked the participants with deciding whether the clip they heard was real or fake.

Binary. We presented 20 randomly chosen clip pairs (labeled 'A' and 'B') comprising the same spoken sentence. Each pair contained a clip uttered by the human speaker and a clip produced by VITS. We randomized the order of the fake and real clips and asked the participants to decide which clip was fake. We included this scenario to see if contextual information helped detection.

Familiarization treatment. In addition to the two configurations, we randomly assigned half of the participants to a familiarization treatment group. We included the treatment to verify the existing literature and understand if humans could be trained to detect deepfakes like an ML model. We showed participants in the treatment group five deepfake utterances before commencing the main detection task. We informed the participants that these examples were synthesized and allowed them to listen to the clips multiple times. These clips were distinct from the stimuli used in the main task.

For participants in the control group, we gave them a filler task. In this task, we asked participants to list potential applications of synthesized speech and to provide their opinion about whether synthesized audio will positively or negatively impact society.

Participants

We recruited participants via the Prolific platform. We filtered for participants fluent in English and Mandarin, as fluency affects detection performance [30]. We paid participants at a rate of £7.25 per hour. To encourage more thoughtful responses, we informed participants they could receive a £1.00 bonus if their detection scores were in the top 50%. Overall, we recruited 529 participants. The mean age was 28.9 years old, and 50.6% identified as male. Table 5.12 contains a more detailed breakdown of the demographics by treatment group.

Table 5.12 Number of participants by group.

Group	English			Mandarin		
	n	*Age (SD)*	*Male (%)*	*n*	*Age (SD)*	*Male (%)*
Unary no familiarization	76	26.8 (8.1)	55.2	65	31.0 (10.4)	44.6
Unary familiarization	65	26.7 (7.3)	56.9	54	31.4 (8.7)	44.4
Binary no familiarization	60	27.5 (7.2)	53.3	70	31.8 (9.0)	48.8
Binary familiarization	80	27.4 (7.3)	57.5	59	29.1 (8.5)	39.6
Overall	281	27.1 (7.5)	55.8	248	30.9 (9.2)	44.5

https://doi.org/10.1371/journal.pone.0285333.t002

Ethics statement

The study was reviewed and exempted by the Department of Security and Crime Science's ethics board at University College London. All participants were notified about the purpose of the study and were over the age of 18. Prior to participating, the participants were asked to tick a series of checkboxes to provide informed written consent.

Benchmarking against automated deepfake detectors

To compare the performance of the human participants to automated methods, we trained two artificial neural networks which specialized in detecting speech deepfakes. The two networks used an LFCC-LCNN architecture [41]. LFCC-LCNNs convert raw audio waveforms into two-dimensional representations. They learn by seeing bona fide and deepfake samples and are rewarded for correctly classifying a sample's authenticity. The ASVspoof 2021 challenge used LFCC-LCNNs as baseline models for spoof detection [9].

Hence, they are a reasonable benchmark for our experiments. For more detail about the top-performing speech deepfake detection architectures, we refer the reader to the article summarizing ASVspoof 2021 [9].

We used two versions for each language:

1. In-domain: We trained the networks using the training split of LJSpeech and CSMSC as bona fide samples and created deepfakes by passing the sentences of the training splits through VITS.
2. Out-of-domain: We trained the Mandarin network with FAD [42], another Mandarin-language dataset. We used the pre-trained ASVspoof network [43] for English-language evaluation.

Artificial neural networks are known to perform well when evaluating against samples similar to those seen during training. However, their performance often drops when encountering different examples [44], even if they are in the same language. These differences can be subtle to a human listener and include changes in the speaker's identity or environment. Therefore, we introduce the out-of-domain version for a fairer comparison with human performance, especially as it is unlikely that the participants in our study recognize the LJSpeech and CSMSC identities.

Results

Overall performance

Fig 5.3 summarizes human performance across all of the different groups. We provide breakdowns of the classification choices in Tables 5.13 and 5.14, which aggregate the English and Mandarin results. We completed the analysis using the SciPy [45] and statsmodels [46] Python packages. For further details, the Supporting information contains results per stimulus.

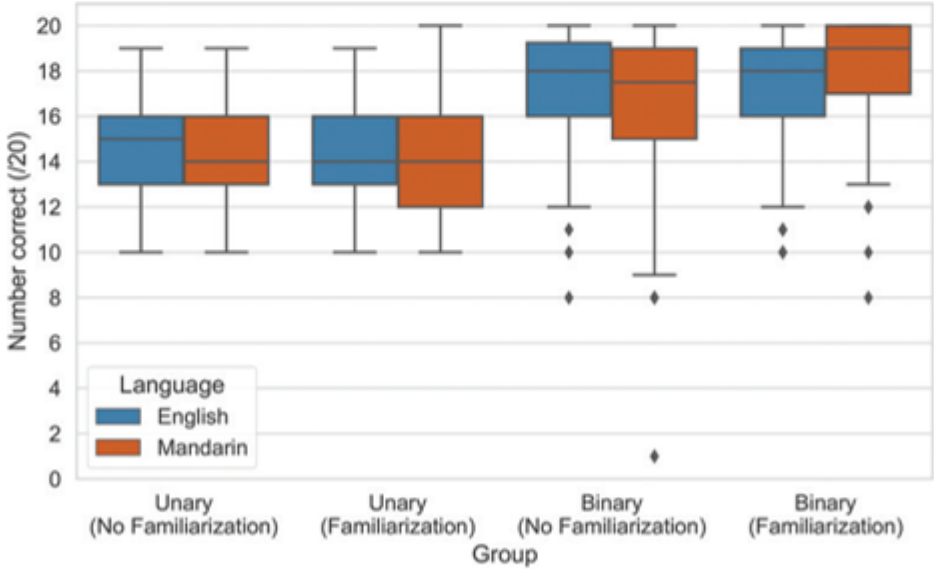


Fig 5.3 Box plot summarizing human performance across the different groups.

https://doi.org/10.1371/journal.pone.0285333.g003

Table 5.13 Confusion matrix for the unary group responses.

		Predicted class	
		Real (2,442)	*Fake (2,678)*
True class	*Real (2,598)*	1,761	837
	Fake (2,522)	681	1,841

$n = 5{,}120$.
Overall accuracy = 70.35%.
Reals correctly identified = 67.78%.
Fakes correctly identified = 73.0%.

https://doi.org/10.1371/journal.pone.0285333.t003

Table 5.14 Confusion matrix for the binary group responses.

		Predicted class	
		Real	*Fake*
True class	*Real*	-	-
	Fake (5,380)	775	4,605

True *real* class labels are not defined in this scenario as participants were asked to choose the fake clip every time. Overall accuracy is equivalent to fakes correctly identified = 85.59%.

https://doi.org/10.1371/journal.pone.0285333.t004

Participants made the correct classifications 70.35% of the time in the unary scenario. They were better at identifying deepfakes (73% accuracy). In comparison, participants correctly identified bona fide examples 67.78% of the time. We speculate the high number of misclassified bona fide samples is partly due to increased skepticism, as participants were aware of the presence of deepfakes through the task briefing. This behavior aligns with observations in Bray et al. [22].

Performance improved under the binary scenario. Participants correctly recognized the deepfake audio in 85.59% of trials. However, the binary setup represents an unrealistic scenario. Even if the speaker's identity is known, reference utterances containing the same speech as the test clip we would like to evaluate are unlikely to be available.

Measuring the effects of interventions

We follow a similar approach to Groh et al. [24] to disentangle the effects of each intervention on performance. We transformed the correct/incorrect results into continuous values by weighting each participant's decision with their provided confidence scores.

The ten-point confidence scale participants completed serves as the mapping function. The lowest score of 0 signals that the participant's choice is a guess, so their confidence in making the right decision corresponds to 50%. In contrast, the highest score of 9 corresponds to 100% belief.

The resulting transformed scores depended on whether the participants made the correct classification. For example, if the participant rated their confidence as 7, this maps to a belief of 88%. If they make the right decision, the adjusted score is 0.88. Conversely, if they make the wrong decision, we subtract the value from 1, resulting in an adjusted score of 0.12.

The revised scores also enable fairer comparisons with the automated deepfake detectors, which output scores between 0 and 1 when evaluating examples. We refer to the revised scores as accuracy scores for the remainder of the text. We also rescale the scores to percentages.

After transforming the results, we analyzed the effects of different interventions on the accuracy scores of participants on each audio clip using linear regression. In addition to language, familiarization and binary intervention, we analyzed the impact of the clip duration. Table 5.15 outlines the results at the overall, unary and binary levels.

Table 5.15 Linear regression results of interventions on confidence-scaled accuracy.

Independent variable	Dependent variable: Confidence-scaled accuracy		
	All (SD)	*Unary (SD)*	*Binary (SD)*
Constant	43.742*** (1.897)	46.394*** (2.791)	71.217*** (2.530)
Mandarin[a]	1.790 (1.404)	1.477 (1.882)	2.152 (2.118)
Familiarization	3.840*** (1.191)	3.758** (1.571)	3.854** (1.802)
Clip length	0.797*** (0.209)	0.375 (0.358)	1.168*** (0.230)
Binary intervention	29.830*** (1.186)	-	-
Observations	10,500	5,120	5,380
R^2	0.165	0.003	0.010
Adjusted R^2	0.165	0.002	0.009
F-Statistic	171.102***	2.455*	11.794***

[a]Dummy variable indicating which language was used in the task. 1 = Mandarin, 0 = English.
*$p < 0.1$.
**$p < 0.05$.
***$p < 0.01$.

https://doi.org/10.1371/journal.pone.0285333.t005

Reference audio helps with deepfake detection. The linear regression results indicate the improvement gained from the binary scenario is statistically significant ($p < 0.001$). Consequently, the results suggest contextual information via reference audio is beneficial for uncovering quirks in synthesized speech.

Training humans to detect deepfakes only helps slightly. The familiarization treatment increases detection accuracy by 3.84% on average ($p = 0.001$). This effect is also present in the unary and binary regression results, improving accuracy by 3.76% ($p = 0.017$) and 3.85% ($p = 0.032$), respectively. However, incorporating familiarizations equates only to an accuracy slightly above chance (52.31%) in the unary setting for the mean clip length (5.76 seconds), ceteris paribus.

It is equally challenging to detect deepfakes in Mandarin and English. Fig 5.3 shows that performance in English and Mandarin is comparable across the different treatment groups. This observation is supported by Table 5.5, which shows Mandarin-speaking participants only outperform their English counterparts by 1.79%, and this effect is not statistically significant ($p = 0.202$).

Shorter speech deepfakes are not easier to identify. As our stimuli varied from 2 to 11 seconds, we included clip length in the regression to verify whether it is easier to discriminate shorter clips. Our results suggest clip length has a negligible impact on accuracy, improving performance by only 0.80% for each additional second. Our scatter plot (Fig 5.4) supports this and shows no relationship between the two variables. These findings conflict with Watson et al. [28], who suggest it is easier to identify shorter deepfakes.

https://doi.org/10.1371/journal.pone.0285333

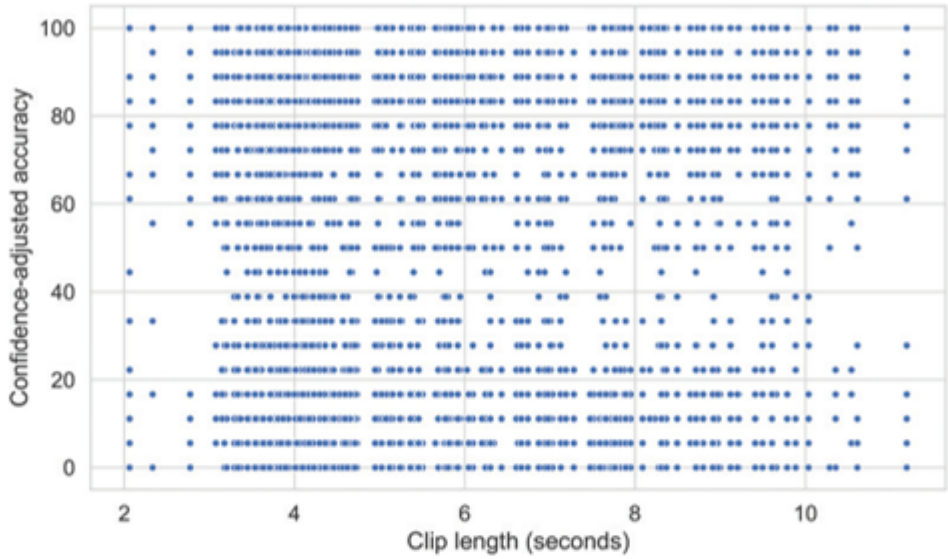


Fig 5.4 Scatter plot showing the relationship between clip length and confidence-scaled accuracy.

https://doi.org/10.1371/journal.pone.0285333.g004

Analyzing performance against time

In addition to analyzing the treatment effects, we examine whether the hypothesis of spending more time on the task improves performance.

Listening to the clips more frequently does not aid detection. We recorded the number of times participants clicked on each audio clip and compared the values to accuracy. As shown in Fig 5.5, there is no relationship between the two variables ($\rho = -0.05$, $p < 0.001$).

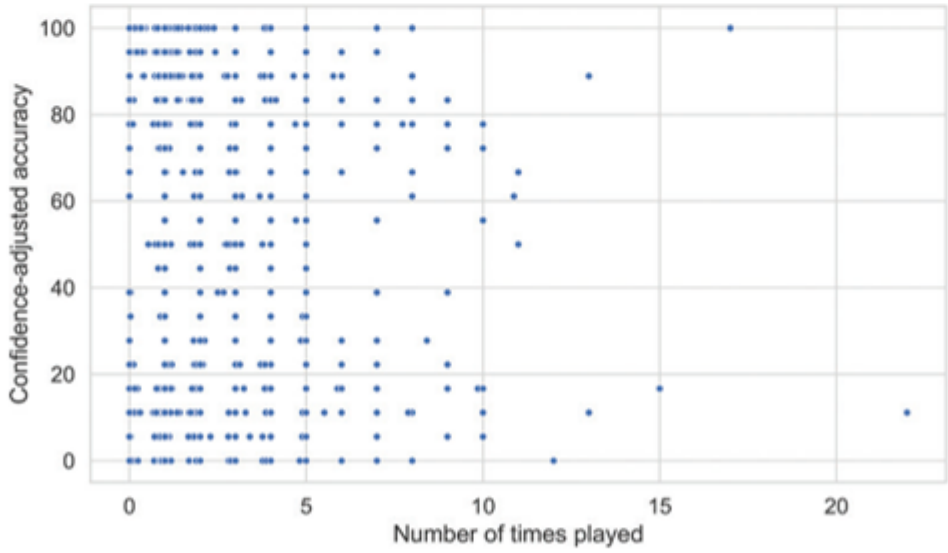


Fig 5.5 Scatter plot showing the relationship between the number of times played and confidence-scaled accuracy.

https://doi.org/10.1371/journal.pone.0285333.g005

Spending more time on the task also does not affect performance. Similar to the above analysis, we compared the time taken to complete the entire task to the total number of clips correctly identified. Fig 5.6 does not indicate a relationship between the two variables ($\rho = 0.10$, $p = 0.018$), suggesting investing more time to complete the task does not improve performance.

Participants do not get better throughout the task without explicit feedback. To understand whether participants improved as they saw more examples and progressed further in the task, we calculated the number of correct responses per question number. If so, we would expect more correct answers in question 20 compared to question 1. Fig 5.7 illustrates the resulting histogram. The histogram shows performance is relatively stable across the questions. This observation indicates participants do not improve throughout the task unless they have explicit feedback, as examined by Groh et al. [24] and Müller et al. [30]. We quantitatively verified the result by conducting a one-way chi-squared hypothesis test against the uniform distribution, which was not statistically significant ($\chi^2 = 6.19$, $p = 0.997$).

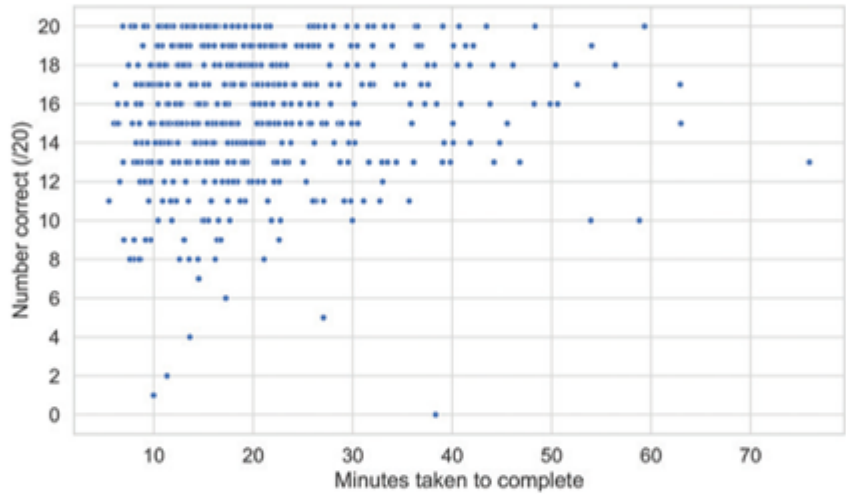


Fig 5.6 Scatter plot showing the relationship between minutes taken to complete and correctness scores.
https://doi.org/10.1371/journal.pone.0285333.g006

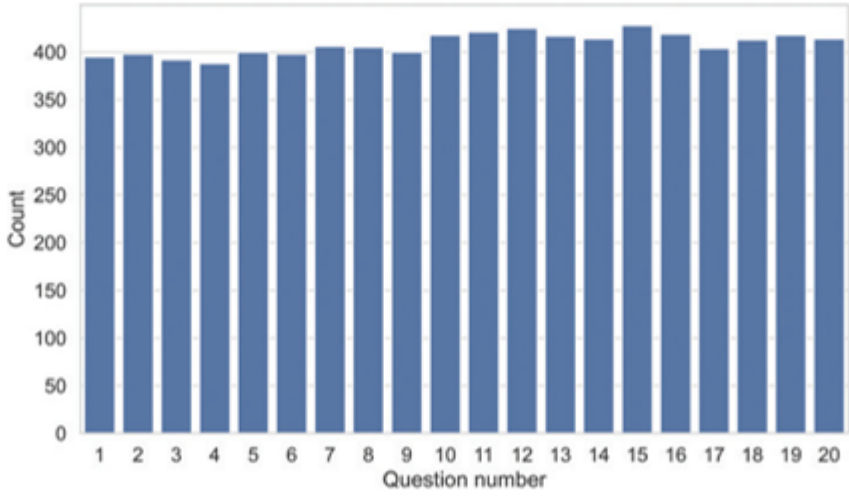


Fig 5.7 Histogram of correct responses across question number.
https://doi.org/10.1371/journal.pone.0285333.g007

Comparing human performance to automated detectors

The following section compares human performance to automated deepfake detectors. For comparability, we use commonly-reported performance metrics found in ML literature.

- **Receiver operating characteristic (ROC)**: These plots represent discriminatory ability. They compare true positive rates against false positive rates at different thresholds.
- **The area under the receiver operating characteristic (AUROC)**: This score summarizes ROCs into a single value. 50% AUROC indicates all predictions are guesses, whereas 100% AUROC means perfect discrimination between bona fides and deepfakes in all trials.
- **Equal error rate (EER)**: This describes the point on ROCs where the true positive and false positive rates are equal.

Fig 5.8 displays the AUROC and EER scores. We include only the unary scenario in this analysis as the inference setup between humans and automated detectors is more comparable. Both evaluate one clip at a time. We aggregated the English and Mandarin results as we observed similar results.

Human performance is less sensitive to unknown conditions compared to automated detectors. The no familiarization (AUROC = 73.83%) and familiarization curves (AUROC = 75.54%) confirm humans performed better than chance. The curves also support the linear regression result. Showing participants examples of deepfakes only had a minute impact on performance. However, performance was quite unreliable: on average, humans incorrectly classified clips a quarter of the time. Humans underperformed the in-domain automated detectors, which had perfect discrimination ability (AUROC = 100% for both languages). However, out-of-domain detectors often incorrectly classified bona fides as deepfakes (AUROC = 25.31%). Based on this behavior, humans are more robust to unknown factors, such as speaker identity.

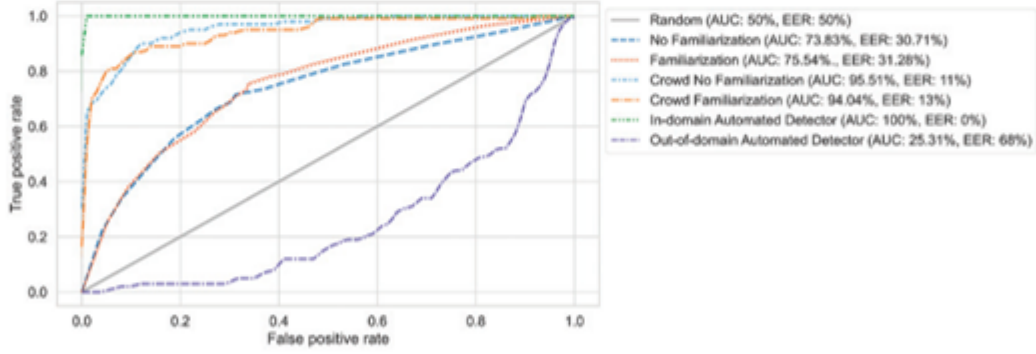


Fig 5.8 Receiver operator curves under the unary scenario.

https://doi.org/10.1371/journal.pone.0285333.g008

Crowd speech deepfake detection is comparable to the top-performing automated detectors. Per Groh et al. [24], we averaged participants' accuracy scores per clip to calculate the crowd-sourced responses. Like the results observed with video stimuli [24], crowd performance is on par with the in-domain detector. However, the benefit of familiarizing participants dissipates when averaging responses. The crowd no familiarization and crowd familiarization AUROCs are similar at 95.51% and 94.04%, respectively.

Freeform text analysis

To understand how participants assessed the genuineness of audio clips, we analyzed their freeform text responses. We grouped responses by language, clip authenticity, and whether participants made the correct choice. We then created word clouds using tf-idf weightings. Tf-idf measures the importance of a word within a document compared to a collection of documents to account for frequently appearing words [47]. Figs 5.9 and 5.10 show the English and Mandarin word clouds.

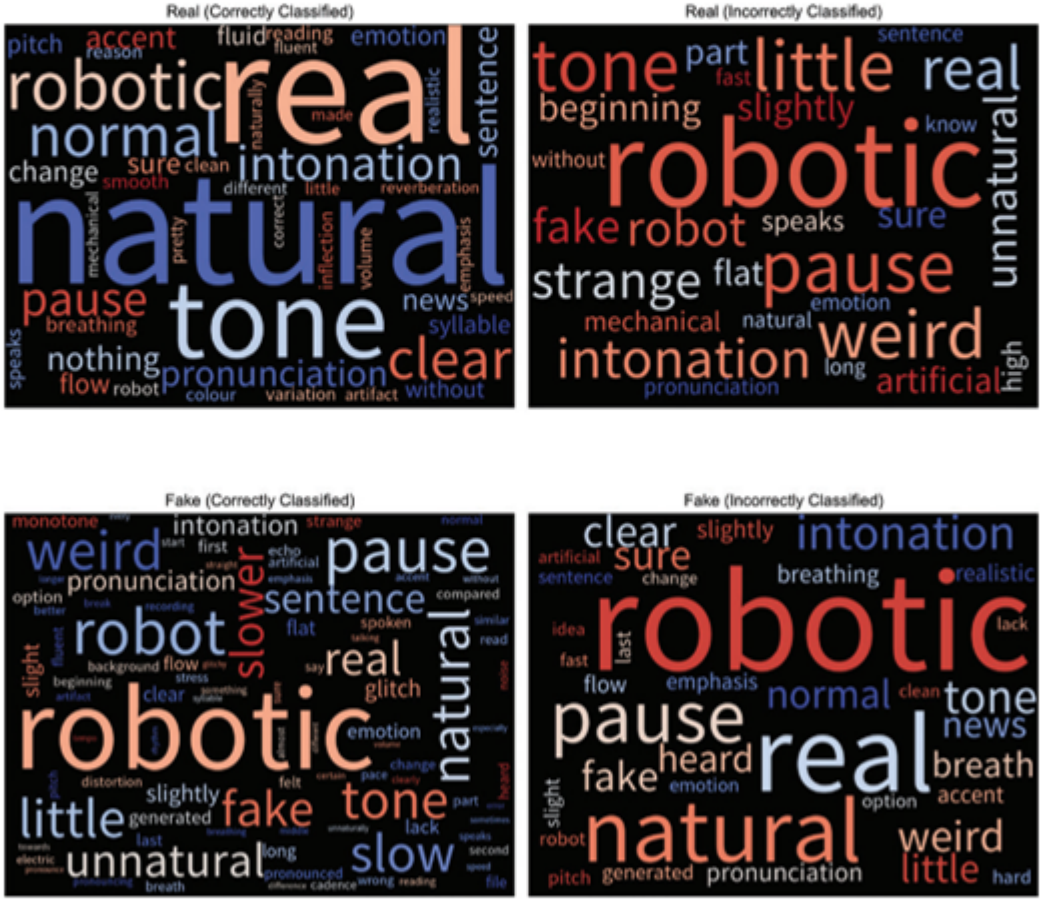


Fig 5.9 Word clouds containing justifications for the English-language clips.

https://doi.org/10.1371/journal.pone.0285333.g009

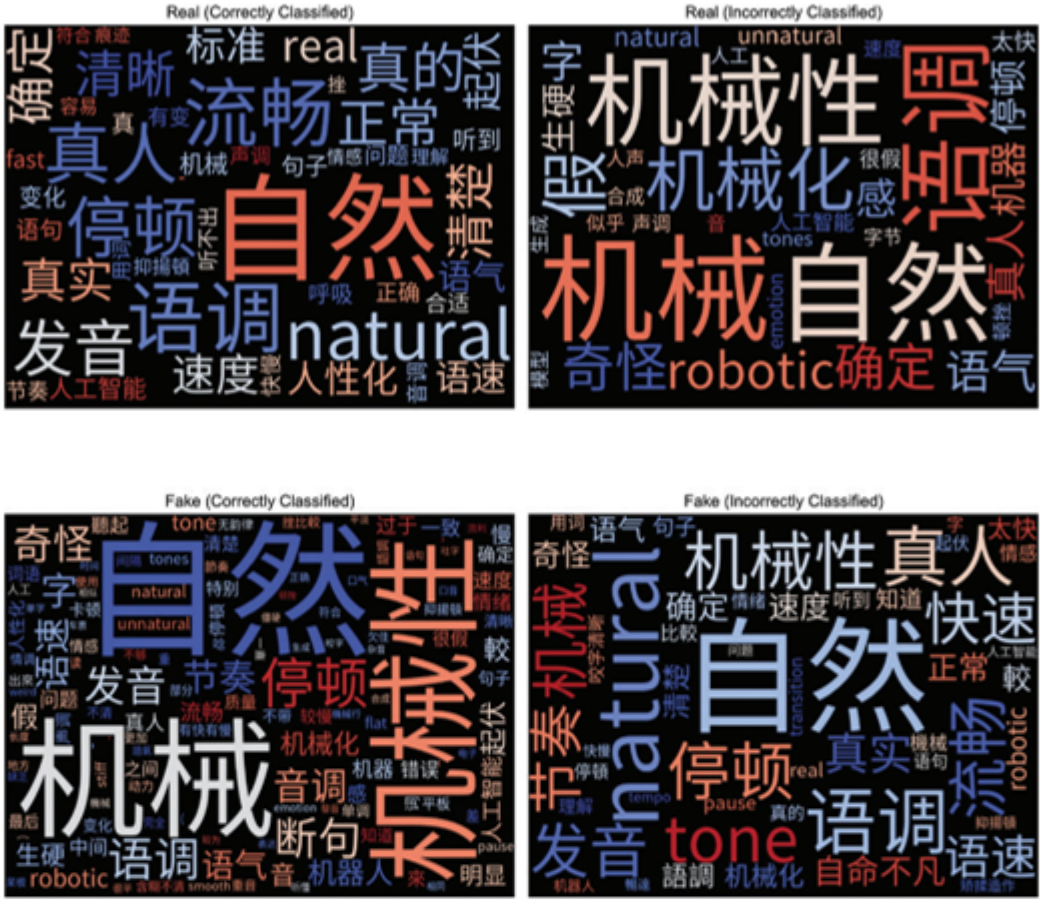


Fig 5.10 Word clouds containing justifications for the Mandarin-language clips. Note participants for the Mandarin tasks provided justifications in both Mandarin and English.

https://doi.org/10.1371/journal.pone.0285333.g010

Participants referred to the same characteristics regardless of whether they made the correct decisions. For example, in Fig 5.9, participants who correctly classified bona fide utterances as legitimate (in the top left of Fig 5.9) mentioned pauses, tone and intonation. However, participants who incorrectly categorized bona fide utterances as fake (top right of Fig 5.9) also referred to these exact attributes. We compared responses by the actual label of the clips and whether participants made the correct response. We did not find substantial differences between these segments. Therefore, automated detectors that incorporate these human characteristics would produce limited improvements. We observed this activity in both English and Mandarin. They tended to rely on intuition to make classifications, referring to naturalness (自然) and robotic (机械) sounds. Beyond intuition, English and Mandarin participants also commonly referenced pauses (停顿), intonation (语调), pronunciation (发音), and speed (速度).

Regarding differences between languages, there were more references to breathing among the English-speaking participants. In contrast, Mandarin-speaking participants mentioned the speaker's cadence (节奏), pacing between words (断句), and fluency (流畅). This result may be due to differences in timing properties between the two languages. English is stress-timed, while Mandarin is syllable-timed [48].

Limitations

Although our setup enabled comparison with automated detectors, it does not necessarily reflect more realistic scenarios where a listener may encounter speech deepfakes.

Firstly, the balance of deepfakes we presented in our experiment does not reflect the proportion that occurs in the wild. Participants were equally likely to encounter deepfakes as bona fides in the task. However, AI-generated content (including the use of deepfakes for nefarious purposes) is still rare for now. In addition, we could expect participants to be much more attentive to the occurrence of deepfakes as we informed them about the nature of the task.

Moreover, we minimized contextual information in our stimuli. For example, we do not examine situations where the listeners' contextual knowledge (such as awareness of the speaker's identity, emotional status, the number of parties in a conversation, or political affiliations) may have informed their decisions. These aspects may be relevant to typical use cases where speech deepfakes may arise, such as false news propagation [5]. Future work could look at exploring how these characteristics influence detection.

Additionally, we asked participants in both languages to listen to utterances purporting to originate from a single female speaker. Given that age and gender of speakers influence speech perception [49, 50], future work could consider how varying speaker identity affects deepfake detection performance.

To generate our deepfake stimuli, we used an older approach which is not necessarily illustrative of the state-of-the-art speech synthesis algorithms. Although our results indicate how well humans can detect speech deepfakes generated with limited-computational resources, they may not faithfully reflect performance under the most current conditions.

Discussion

Humans can detect speech deepfakes, but not consistently. They tend to rely on naturalness to identify deepfakes regardless of language. As speech synthesis algorithms improve and become more natural, it will become more difficult for humans to catch speech deepfakes.

Although there are some differences in the features that English and Mandarin speakers use to detect deepfakes, the two groups share many similarities. Therefore, the threat potential of speech deepfakes is consistent despite the language involved.

It will be easier for adversaries to generate more deepfakes as the computational barrier for synthesizing data lowers. More deepfakes in the wild will have a knock-on effect. Adversaries will have more opportunities to scale their operations, particularly for disinformation such as impersonations and spear phishing [6].

Ultimately, the battle between deepfake creation and detection is an arms race [51]. How can we defend against falling prey to deepfake trickery? Our binary scenario shows that comparing against reference audio is helpful if we know the speaker's identity. However, we do not always have this information.

Increasing awareness by showing people examples of deepfake audio has a limited effect, as demonstrated by our familiarization results. Spending more time evaluating the clips does not seem to help either.

To summarize, attempting to improve human detection capabilities is unrealistic. We show that even in a controlled environment where the task is easier (participants are aware of the presence of speech deepfakes and the deepfakes are not created using state-of-the-art speech synthesizers), deepfake detection is not high.

https://doi.org/10.1371/journal.pone.0285333

Our results suggest the need for automated detectors to mitigate a human listener's weaknesses. Automated detectors' performance on in-domain data indicates they can pick up on subtleties that humans cannot. However, we show they are brittle and fail to work when there are changes in the test audio's environmental conditions. Given the extent of human limitations and the increasing availability of computational resources for deploying detectors, research should focus on improving these detectors. In the meantime, crowd-sourcing is a reasonable mitigation. We confirm crowd performance is on par with the top-performing automated detectors and is not as brittle. Extending fact-checking tools to include audio evaluations is one way to protect against deepfake threats.

Supporting Information

S1 Fig. Confidence-adjusted accuracy scores per clip (English, unary, no familiarization).
(TIF)

S2 Fig. Confidence-adjusted accuracy scores per clip (English, unary, familiarization).
(TIF)

S3 Fig. Confidence-adjusted accuracy scores per clip (English, binary, no familiarization).
(TIF)

S4 Fig. Confidence-adjusted accuracy scores per clip (English, binary, familiarization).
(TIF)

S5 Fig. Confidence-adjusted accuracy scores per clip (Mandarin, unary, no familiarization).
(TIF)

S6 Fig. Confidence-adjusted accuracy scores per clip (Mandarin, unary, familiarization).
(TIF)

S7 Fig. Confidence-adjusted accuracy scores per clip (Mandarin, binary, no familiarization).
(TIF)

S8 Fig. Confidence-adjusted accuracy scores per clip (Mandarin, binary, familiarization).
(TIF)

Acknowledgments

Kimberly T. Mai thanks Kelvin Ma for assistance with translating the survey materials.

Author Contributions

Conceptualization: Kimberly T. Mai, Sergi Bray, Toby Davies, Lewis D. Griffin.
Data curation: Kimberly T. Mai.
Formal analysis: Kimberly T. Mai.
Investigation: Kimberly T. Mai.
Methodology: Kimberly T. Mai, Sergi Bray, Lewis D. Griffin.
Software: Kimberly T. Mai, Sergi Bray.
Supervision: Toby Davies, Lewis D. Griffin.
Visualization: Kimberly T. Mai.

Writing – original draft: Kimberly T. Mai.
Writing – review & editing: Toby Davies, Lewis D. Griffin.

References

1. Brewster T. Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find. 2021 Oct 14 [Cited 2023 Jan 19]. Available from: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=7dfbccf67559.
2. Stupp C. Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case. 2019 Aug 30 [Cited 2023 Jan 19]. Available from: https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.
3. Choi S, Han S, Kim D, Ha S. Attentron: Few-Shot Text-to-Speech Utilizing Attention-Based Variable-Length Embedding. In: Meng H, Xu B, Zheng TF, editors. Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020. ISCA; 2020. p. 2007–2011.
4. Alspach K. Does your boss sound a little funny? it might be an audio deepfake; 2022 Aug 18 [Cited 2023 Jan 19]. Available from: https://www.protocol.com/enterprise/deepfake-voice-cyberattack-ai-audio.
5. Caldwell M, Andrews J, Tanay T, Griffin L. AI-enabled future crime. Crime Science. 2020;9(1): 1–13. https://doi.org/10.1186/s40163-020-00123-8
6. Mirsky Y, Demontis A, Kotak J, Shankar R, Gelei D, Yang L, et al. The threat of offensive ai to organizations. Computers & Security. 2022; p. 103006.
7. Wu Z, Yamagishi J, Kinnunen T, Hanilçi C, Sahidullah M, Sizov A, et al. ASVspoof: The Automatic Speaker Verification Spoofing and Countermeasures Challenge. IEEE Journal of Selected Topics in Signal Processing. 2017;11(4):588–604. https://doi.org/10.1109/JSTSP.2017.2671435
8. Nautsch A, Wang X, Evans NWD, Kinnunen TH, Vestman V, Todisco M, et al. ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2021;3(2):252–265. https://doi.org/10.1109/TBIOM.2021.3059479
9. Yamagishi J, Wang X, Todisco M, Sahidullah M, Patino J, Nautsch A, et al. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. arXiv preprint arXiv:210900537. 2021 Sep 1 [Cited 2023 Jan 19].
10. Gamage D, Chen J, Ghasiya P, Sasahara K. Deepfakes and Society: What lies ahead? In: Frontiers in Fake Media Generation and Detection. Springer; 2022. p. 3–43.
11. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.
12. Zhang Y, Tiňo P, Leonardis A, Tang K. A survey on neural network interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence. 2021. https://doi.org/10.1109/TETCI.2021.3100641
13. van Huijstee M, van Boheemen P, Das D, Nierling L, Jahnel J, Karaboga M, et al. Tackling Deepfakes in European Policy. European Parliament; 2021. Available from: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2021)690039.
14. Schick N. Deep Fakes and the Infocalypse: What You Urgently Need To Know. Hachette; 2020.

15. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Nets. In: Advances in Neural Information Processing Systems; 2014. Available from: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

16. Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In: Proceedings of the 32nd International Conference on Machine Learning. Available from: https://proceedings.mlr.press/v37/sohl-dickstein15.html

17. Damer N, Saladié AM, Braun A, Kuijper A. MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS).

18. Bitouk D, Kumar N, Dhillon S, Belhumeur P, Nayar SK. Face Swapping: Automatically Replacing Faces in Photographs ACM Trans Graph. 2008;27(3):1–8.

19. Zhang T. Deepfake generation and detection, a survey. Multimedia Tools and Applications. 2022;81(5):6259–6276. https://doi.org/10.1007/s11042-021-11733-y

20. Barari S, Lucas C, Munger K. Political Deepfakes Are As Credible As Other Fake Media And (Sometimes) Real Media [Preprint]; 2021 [Cited 2023 Jan 19]. Available from: osf.io/cdfh3.

21. Appel M, Prietzel F. The detection of political deepfakes. Journal of Computer-Mediated Communication. 2022;27(4). https://doi.org/10.1093/jcmc/zmac008

22. Bray SD, Johnson SD, Kleinberg B. Testing Human Ability To Detect Deepfake Images of Human Faces. arXiv preprint arXiv:221205056. 2022 Dec 7 [Cited 2023 Jan 19].

23. Nightingale SJ, Farid H. AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proceedings of the National Academy of Sciences. 2022;119(8):e2120481119. https://doi.org/10.1073/pnas.2120481119 PMID:35165187

24. Groh M, Epstein Z, Firestone C, Picard R. Deepfake detection by human crowds, machines, and machine-informed crowds. Proceedings of the National Academy of Sciences. 2021;119(1):e2110013119. https://doi.org/10.1073/pnas.2110013119

25. Köbis NC, Doležalová B, Soraperra I. Fooled twice: People cannot detect deepfakes but think they can. iScience. 2021;24(11):103364. https://doi.org/10.1016/j.isci.2021.103364 PMID:34820608

26. Tahir R, Batool B, Jamshed H, Jameel M, Anwar M, Ahmed F, et al. Seeing is believing: Exploring perceptual differences in deepfake videos. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; 2021. p. 1–16.

27. Geirhos T, Zimmerman RS, Bilodeau BL, Brendal W, Kim, B. Don't trust your eyes: on the (un) reliability of feature visualizations. arXiv preprint arXiv: 230604719. 2023 Jun 13 [Cited 2023 Jun 13].

28. Watson G, Khanjani Z, Janeja VP. Audio Deepfake Perceptions in College Going Populations. arXiv preprint arXiv:211203351. 2021 Dec 6 [Cited 2023 Jan 19].

29. Wang X, Yamagishi J, Todisco M, Delgado H, Nautsch A, Evans N, et al. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. Computer Speech & Language. 2020;64:101114. https://doi.org/10.1016/j.csl.2020.101114

30. Müller NM, Pizzi K, Williams J. Human perception of audio deepfakes. In: Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia; 2022. p. 85–91.

31. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 8110–8119.

32. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, et al. The Deepfake Detection Challenge Dataset. arXiv preprint arXiv:200607397. [Cited 2023 Mar 27]. Dataset.

33. Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 3207-3216.

34. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 1–11.

35. Perov I, Gao D, Chervoniy N, Liu K, Marangonda S, Umé C, et al DeepFaceLab: Integrated, flexible and extensible face-swapping framework. arXiv preprint arXiv:2005.05535. [Cited 2023 Mar 27]. Dataset.

36. Kumar K, Kumar R, De Boissiere T, Gestin L, Teoh WZ, Sotelo J, et al. MelGAN: Generative adversarial networks for conditional waveform synthesis. In: Proceedings of Advances in Neural Information Processing Systems; 2019;32.

37. Ito K, Johnson L. The LJ Speech Dataset; 2017 [Cited 2023 Jan 19]. Dataset. https://keithito.com/LJ-Speech-Dataset/.

38. Databaker. Chinese Standard Mandarin Speech Corpus; 2019 [Cited 2023 Jan 19]. Dataset. https://www.data-baker.com/open_source.html.

39. Watanabe S, Hori T, Karita S, Hayashi T, Nishitoba J, Unno Y, et al. ESPnet: End-to-End Speech Processing Toolkit. In: Proceedings of Interspeech; 2018. p. 2207–2211.

40. Kim J, Kong J, Son J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: International Conference on Machine Learning. PMLR; 2021. p. 5530–5540. Available from: https://proceedings.mlr.press/v139/kim21f.html.

41. Wang X, Yamagishi J. A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection. In: Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August—3 September 2021. ISCA; 2021. p. 4259–4263.

42. Ma H, Yi J, Wang C, Yan X, Tao J, Wang T, et al. FAD: A Chinese Dataset for Fake Audio Detection. arXiv preprint arXiv:220712308. 2022 Jul 12. [Cited 2023 Jan 19].

43. Delgado H, Evans N, Kinnunen T, Lee KA, Liu X, Nautsch A, et al. ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. arXiv preprint arXiv:210900535. 2021 Sep 1 [Cited 2023 Jan 19].

44. Shen Z, Liu J, He Y, Zhang X, Xu R, Yu H, et al. Towards out-of-distribution generalization: A survey. arXiv preprint arXiv:210813624. 2021 Aug 31. [Cited 2023 Jan 19].

45. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020;17:261–272. https://doi.org/10.1038/s41592-019-0686-2 PMID:32015543

46. Seabold S, Perktold J. statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference; 2010.

47. Leskovec J, Rajaraman A, Ullman JD. Mining of Massive Datasets. 2nd ed. USA: Cambridge University Press; 2014.

48. Mok PPK, Dellwo V. Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. In: Proc. Speech Prosody 2008; 2008. p. 423–426.

49. Hummert ML, Shaner JL, Garstka TA, Henry C. Communication with older adults: The influence of age stereotypes, context, and communicator age. Human communication research. 1998;25(1):124–151. https://doi.org/10.1111/j.1468-2958.1998.tb00439.x

50. Strand EA. Uncovering the role of gender stereotypes in speech perception. Journal of language and social psychology. 1999;18(1):86–100. https://doi.org/10.1177/0261927X99018001006

51. Chesney R, Citron D. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. Foreign Affairs. 2019 Jan/Feb;98:147.

5.9 Discussion Topics

For this chapter, we are to discuss types of cybercrime and their impact on society. The rapid growth of technology and internet usage have brought many advantages to society. But they have also introduced a new form of criminals that use the internet to engage in various types of cybercrimes. These are crimes that are committed through the internet and they are quite common and have serious consequences to the victims. Here are some potential discussion points that can be used to develop these topics further:

1. How have ransomware attacks changed in recent times and which industries have been the most vulnerable?
2. How do phishing and other social engineering techniques use human weaknesses to carry out crimes?
3. Are cybercriminal networks as structured and advanced as organized crime groups?
4. How is cybercrime prevalent in countries with low and high levels of technological advancement?
5. In addition to financial loss what are the effects of identity theft on the psychological and social well being of victims in the long run?
6. Can artificial intelligence systems substitute human security analysts or is a team effort necessary?
7. How successful are biometric methods (such as facial recognition, fingerprints, and iris scans) in controlling cybercrime and what are the dangers?
8. Which is more effective in fighting cybercrime: prevention or reaction? Do organizations need to focus on stopping cyber attacks or helping the victims after an attack?
9. The use of honeypots and intrusion detection systems: How do these tools help in detecting and identifying cybercriminals?
10. Will current encryption methods become obsolete with the advancement of quantum computing and if so what other methods are there?
11. At what level does cybersecurity become an invasion of privacy and how should we trade off personal privacy for the protection of the group?
12. Should there be more legal immunity for ethical hackers (white-hats) and at what point does ethical hacking turn into criminal activity?
13. How much should governments and companies be able to monitor people to prevent cybercrime?
14. What are the possible ethical implications of using AI to detect, forecast and fight cybercrimes? How can biases be prevented in AI algorithms?
15. Should cybersecurity professionals be involved in the fight against online misinformation and what ethical considerations should they follow?
16. Can cybersecurity professionals help in combating online misinformation and if so, what ethical considerations should they bear in mind?
17. How can law enforcement agencies from different countries collaborate in order to prevent cybercrime and what problems may arise during international cooperation?
18. What kind of problems do law enforcement agencies encounter while prosecuting cybercriminals especially when the evidence is digital in nature?

19. How important is digital forensics in creating credible legal cases against cybercriminals?
20. Should law enforcement agencies be the ones to fight against cybercrime or should the private sector take the lead in developing cybersecurity measures?
21. How can law enforcement agencies fight against cybercrime when they have limited resources and the technology is advancing quickly?

Use the information gained from this book, along with your personal experiences, thoughts, and judgments to enhance your post and encourage further discussion.

5.10 Projects

5.10.1 Cybersecurity Breach at the CIA: Ethical and Societal Implications of Information Disclosure

Abstract: This project aims at a detailed analysis of the moral, security and privacy implications of the unauthorized access and leak of a U.S. Central Intelligence Agency official's private AOL account details. Through a detailed analysis of the benefits and risks of such exposures, the project will be able to evaluate the morality of leaking the information by WikiLeaks. The case study approach, together with ethical principles and the analysis of cybersecurity issues, will be used to study this challenging topic.

Research Questions: Are there any benefits in publishing the details of a CIA official's hacked AOL account for public use? What are the risks and problems involved in releasing such information from the perspective of cybersecurity and privacy? Under what ethical frameworks can the release of the hacked information be judged, and what are the consequences of such a judgment? How does the exposure of sensitive data conform with the laws of the land and the general public's views on privacy and security?

Methodology: Case Study Analysis: A detailed look at the situation including the sequence of events and what was included in the leaked information. Interviews with experts in cybersecurity, ethics and law will be conducted to capture the various impacts of the case. Ethical Framework Application: A utilitarian, deontological and virtue ethical approach will be used to evaluate the morality of the information leak. Risk-Benefit Analysis: The advantages and disadvantages of making such information public from a security point of view.

Expected Outcomes: A report that provides an analysis of the advantages and disadvantages of the WikiLeaks leak and also touches on ethical, legal and cybersecurity aspects. A thorough ethical consideration to determine if the leak was warranted. A set of recommendations for future policies and security measures to avert such cases and protocols for handling such cases. An improved perception of the relationship between the openness and protection of information in the modern era.

Timeline:

- Week 1: Project Orientation and Preliminary Research. Define the project's approach and determine the required resources. Initial research on the case study and preliminary findings on cybersecurity breaches started.
- Week 2: Interviews and Ethical Framework Selection. Interviews with the experts will be done to gain different viewpoints. Ethical frameworks will be chosen and studied for analysis.
- Week 3: Data Analysis and Ethical Evaluation. Data from the research will be evaluated and ethical principles will be used to judge the situation. The focus will be on the ethicality of the publication as the report is being written.

- Week 4: Finalization and Presentation. The report will be completed with a risk-benefit assessment and policy suggestions. A presentation will be made to present the results and conclusions to fellow students and faculty.

Conclusion: Therefore, by the end of the project, participants will have encountered some of the current issues at the nexus of cybersecurity, ethics, and public policy, and, as a result, they will have a better grasp of the real-life implications of data privacy and security breaches.

5.10.2 Unauthorized Access for Authorized Safety: Evaluating the Role of Hackers in Cybersecurity

Abstract: This four-week undergraduate project investigates the problematic view that hackers assist cybersecurity systems by revealing vulnerabilities and making them public. This session investigates how hackers' security threats prevent numerous security weaknesses from being both discovered and fixed. The research explores the idea of hackers as unintentional digital champions who help security by revealing weaknesses to the public but questions whether their activities create genuine threats to privacy and online protection.

Research Questions: Hacking activities without authorization serve as valid means to improve cybersecurity by revealing system weaknesses. How much value do hackers bring to security improvements through their actions when compared to traditional authorized penetration tests? Legal and ethical standards define hacker conduct but how these boundaries influence public opinion about their cybersecurity roles remains uncertain.

Methodology: Literature Review: A comprehensive review of existing literature about hacking effects on cybersecurity development will be performed. Cybersecurity experts together with ethical hackers and legal professionals must be interviewed to obtain their insights regarding unauthorized hacking advantages and disadvantages. Research multiple documented hacking incidents that resulted in major security improvements and evaluate their success compared to standard security audit results. The ethical nature of hacker activities will be examined through ethical frameworks that compare their benefits for the public with their capacity to cause damage.

Expected Outcomes: The report presents an objective evaluation of hacker involvement in cybersecurity practices. A comparative analysis of security improvements resulting from unauthorized hacking versus formal penetration testing. Expert opinions and ethical reasoning support ethical conclusions about hacking activities in cybersecurity. The report will outline how the cybersecurity community along with legal systems should maximize hacking benefits while reducing associated dangers.

Timeline:

- Week 1: Project Planning and Literature Review. The project objectives should be finalized and research on hacking and cybersecurity should begin. A structure for evaluating ethical hacking activities needs to be developed.
- Week 2: Expert Interviews and Case Study Selection. A selection of expert participants from different sectors needs to be interviewed. Choose specific hacking incidents that led to better security as the subject of study.
- Week 3: Data Analysis and Ethical Considerations. The collected interview and case study data need to be analyzed. The actions of hackers and their resulting outcomes will be evaluated through ethical theory application.

- Week 4: Synthesis and Reporting. The project combines insights from literature reviews with interview data and case studies as well as ethical evaluation results. The report will conclude with a detailed assessment of hacker involvement in cybersecurity by presenting the complex relationship between hacking activities and system protection. The main findings along with suggested actions should be presented through either a presentation or an infographic.

Conclusion: The research project establishes a deep understanding of hackers' multifaceted position in cybersecurity by separating criminal purposes from beneficial effects. The research will provide guidance on utilizing hacker expertise to defend digital assets in a manner that maintains legal and ethical standards.

5.11 Exercises

1. How does cyberbullying differ from traditional bullying?
2. What is a data breach, and what can cybercriminals do with stolen data?
3. What is cyberterrorism, and what types of targets might cyberterrorists attack?
4. How does cybercrime impact individuals financially?
5. What are the psychological effects of cyberbullying on victims?
6. Why is it difficult for law enforcement to prosecute cybercriminals?
7. How does cybercrime undermine trust in online transactions?
8. What are the societal challenges posed by cybercrime?
9. What role do firewalls play in preventing cybercrime?
10. Why is it important to regularly update software and systems?
11. How does encryption help protect data?
12. What is the purpose of conducting regular risk assessments in cybersecurity?
13. How can education and awareness programs prevent cybercrime?
14. What is an incident response plan, and why is it important?
15. What are the benefits of using AI in cybercrime detection?
16. Why is collaboration between the public and private sectors important in combating cybercrime?
17. What is vulnerability management, and how does it help prevent cybercrime?
18. Why is privacy protection an important ethical consideration in cybersecurity?
19. What is responsible disclosure in the context of cybersecurity?
20. Why should organizations be transparent about their cybersecurity practices? security measures.
21. What is the ethical dilemma in balancing security with personal freedoms?
22. What ethical concerns arise from the use of offensive cyber capabilities?
23. How does transparency promote accountability in cybersecurity?
24. What is the challenge of maintaining both privacy and security in cybersecurity?

25. What is the primary role of law enforcement agencies in preventing cybercrime?
26. Why is international cooperation essential in fighting cybercrime?
27. What challenges do law enforcement agencies face when investigating cybercrime?
28. What is the ethical debate surrounding encryption backdoors for law enforcement?
29. Why do law enforcement agencies advocate for encryption backdoors?

Chapter

6

Intellectual Property

- The history of intellectual property law and its application to computer technology
- Examples of copyright, patent, and trademark infringement in the digital age
- The Importance of Fair Use in Copyright Law
- Significant fair-use cases and their impacts
- The Digital Millennium Copyright Act
- The impact of open-source software and the sharing economy on intellectual property rights
- The role of government in protecting intellectual property

6.1 The Evolution of Intellectual Property Law and Its Application to Computer Technology

Intellectual property law serves as a protective mechanism which grants exclusive rights to creators and inventors for their original works. The quick progress of computer technology during the past forty years has created new problems and prospects for intellectual property protection. The presentation tracks the intellectual property law development through time and shows how it affects computer technology by identifying vital historical stages and governing principles of this active field [91, 52, 201, 138, 84].

1. The earliest roots of intellectual property law emerged in ancient societies when creative producers received specific privileges and rights for their works. Intellectual property concepts did not emerge until the Renaissance period. The Statute of Anne in 1710 became known as the initial copyright law which created copyright protection standards that established exclusive rights for creators.
2. Computer technology introduced new intellectual property law issues which emerged during its development. The increasing use of computers made it necessary to protect software alongside technological inventions. The United States Patent Act of 1790 represented a critical development which provided inventors with exclusive rights to protect their inventions. Court decisions starting in the 1980s established software as patentable subject matter through cases including Diamond v. Diehr (1981) and Diamond v. Chakrabarty (1980).
3. The digital revolution forced copyright law to confront new problems which it had never before encountered. The digital nature of information enabled fast duplication and distribution which threatened creators' control over their works along with their ability to receive compensation. Through the Digital Millennium Copyright Act (DMCA) of 1998 the United States sought to address copyright challenges by making technological copyright protection measures subject to criminal prosecution.
4. Software positions itself as a special case within intellectual property legal frameworks. The protection of software rests between copyright and patent laws according to its distinctive features. Under copyright law software developers receive protection for their work's expression and under

patent law developers gain protection for software processes and functionality. The debate about software patents continues to expand because experts seek a proper approach to balance innovation promotion with monopoly prevention.

5. The open-source movement rejected conventional intellectual property approaches by supporting uncontrolled software distribution and collaborative development activities. The GNU General Public License (GPL) which Richard Stallman launched in 1989 became a well-known example of the open-source movement. Through this license users gain the freedom to use software while maintaining specific rights and obligations for distribution and modification purposes.
6. The need to synchronize intellectual property laws across nations has grown more crucial because technology operates without international borders. The World Intellectual Property Organization (WIPO) Copyright Treaty together with the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) works to create standardized intellectual property laws and enforcement systems which enable worldwide innovation protection for creators.

In conclusion, intellectual property law continues to require updates in its application to computer technology because of the need to match legal systems to technological developments. Intellectual property law faces an ongoing challenge to maintain proper creator protection while allowing innovation to thrive because of computer technology's rapid evolution. The digital era requires societies to understand and apply copyright and patent laws to create an environment which stimulates creativity while motivating inventions and supporting technological growth.

6.2 Examples of Copyright, Patent, and Trademark Infringement in the Digital Age

The digital age has brought about a multitude of opportunities for sharing and accessing creative works, technological innovations, and brand identities. However, with these advancements, instances of copyright, patent, and trademark infringement have become more prevalent. This session explores examples of infringement in the digital age, highlighting the challenges faced by creators, inventors, and brand owners in protecting their intellectual property.

1. Copyright Infringement: Copyright infringement occurs when someone reproduces, distributes, performs, displays, or creates derivative works of copyrighted material without authorization. In the digital age, online platforms and peer-to-peer networks have facilitated the widespread sharing of copyrighted content, leading to numerous instances of infringement. Some examples include:
   (a) Online Piracy: Websites and services that provide access to unauthorized copies of movies, music, books, and software, such as torrent sites or streaming platforms that offer copyrighted content without proper licensing.
   (b) Plagiarism: Copying and presenting someone else's written work, such as articles, blog posts, or academic papers, as one's own without permission or attribution.
   (c) Unauthorized Remixes and Mashups: Creating new works by combining copyrighted material without obtaining the necessary licenses or permissions from the original creators.
2. Patent Infringement: Patent infringement occurs when someone makes, uses, sells, or imports a patented invention without permission from the patent holder. In the digital age, patent infringement is often associated with technological inventions and software. Some examples include:

(a) Software Infringement: Developing and distributing software that incorporates patented algorithms, processes, or methods without obtaining a license or authorization from the patent holder.

(b) Patent Trolls: Entities that acquire patents solely for the purpose of filing lawsuits against alleged infringers, often targeting technology companies with broad or vague patents.

(c) Reverse Engineering: Disassembling and analyzing a patented product or technology to replicate its functionality or design without permission, potentially leading to infringement.

3. Trademark Infringement: Trademark infringement occurs when someone uses a registered trademark or a similar mark in commerce without the permission of the trademark owner, leading to confusion among consumers. In the digital age, trademark infringement has become prevalent due to the ease of online advertising and the proliferation of e-commerce. Some examples include:

(a) Cybersquatting: Registering, trafficking, or using domain names that are identical or confusingly similar to established trademarks, with the intention of profiting from the reputation of the original brand.

(b) Counterfeit Goods: Producing and selling unauthorized replicas or imitations of branded products, often through online marketplaces, deceiving consumers into believing they are purchasing genuine items.

(c) Keyword Advertising: Using another company's trademarked terms as keywords to trigger online advertisements, potentially causing consumer confusion or dilution of the original brand.

In conclusion, the digital age has presented numerous challenges for intellectual property rights holders, as copyright, patent, and trademark infringement have become more prevalent in online environments. From online piracy and plagiarism to patent trolls and trademark counterfeiting, creators, inventors, and brand owners face ongoing battles to protect their intellectual property. It is essential for legal frameworks, enforcement agencies, and digital platforms to work together to address these infringements, strike a balance between innovation and protection, and safeguard the rights and interests of intellectual property owners in the digital landscape.

6.3 The Importance of Fair Use in Copyright Law

The fair use doctrine stands as a fundamental part of copyright law which enables restricted use of copyrighted materials without needing copyright holder permission. The doctrine serves as a vital mechanism to support both free expression rights and educational activities while advancing arts and scientific progress. The United States implements this doctrine through Title 17 of the U.S. Code while other countries maintain analogous concepts.

The fundamental principle behind fair use exists to maintain equilibrium between copyright holder rights and public authorization to use copyrighted content reasonably. The use does not qualify as infringement because it falls under fair use provisions. Each fair use determination requires individual assessment based on four specific statutory elements.

The first factor analyzes the use's purpose together with its nonprofit educational characteristics. Transformative uses that incorporate new expression or meaning or messages into the original work receive stronger support for fair use status.

Copyrighted works that are published and factual and nonfiction based have better chances of meeting fair use criteria than unpublished and creative works like fiction and music and movies.

The evaluation assesses both the quantitative aspect and the qualitative aspect of the copyrighted material that gets utilized. The quantity of material used is not the sole factor in determining fair use because taking the essential elements of a work negates fair use protection regardless of the quantity.

The use will be considered less fair when it demonstrates a major impact on the sales or market potential of the copyrighted work. The factor protects the copyright owner from economic losses that occur from the use of their material.

The fair use provisions allow people to perform activities including criticism and comment as well as news reporting and teaching alongside scholarship and research. Through fair use provisions educators can teach with copyrighted materials and journalists can report news by using copyrighted content and artists gain permission to produce parodies and commentaries about existing works. The doctrine helps digital humanities research by enabling the necessary access to large corpora of copyrighted texts for text and data mining operations.

The defense of fair use remains incomplete because courts need to make personal determinations that frequently create legal controversies. Legal professionals advise numerous institutions and individuals about fair use because they want to avoid copyright infringement risks.

In conclusion, fair use continues to evolve as technology advances because it affects how we handle copyrighted materials. The internet has introduced new questions regarding fair use boundaries which resulted in important legal cases that define this doctrine's interpretation. The doctrine faces a challenge in adapting to the digital era while preserving its essential mission of supporting creative activities and information sharing.

6.4 Significant Fair-Use Cases and Their Impacts

Fair use cases have become fundamental for defining copyright laws and determining how the fair use doctrine works. These court decisions influence both the involved parties and produce extensive effects on copyright practices and legal interpretations together with creative freedom. The following section examines important cases which significantly affected copyright laws.

- Authors Guild v. Google (2015): The Authors Guild brought forward their lawsuit against Google because the company digitized books without authorization for its Google Books project. The court established that Google's indexing function which enabled users to locate books remained transformative because it created a search system that did not allow users to obtain books without payment. This court decision strongly supported transformative use while impacting the future development of digital libraries and search engines.
- The Supreme Court made a decision in Campbell v. Acuff-Rose Music, Inc. (1994) regarding whether 2 Live Crew could parody Roy Orbison's "Oh, Pretty Woman." The court determined that commercial parodies can be legitimate under fair use provisions when they offer new meanings or expressions. The case established a crucial precedent because it specifically validated parody and satirical content as fair use.
- In the "Betamax case" Sony Corp, of America v. Universal City Studios, Inc. (1984) the court determined that personal time-shifting of complete TV shows qualified as fair use. The court decision shaped both the home video sector and subsequent advancements in personal digital recording systems.
- Harper & Row Publishers, Inc. v. Nation Enterprises (1985): The Nation magazine released portions from Gerald Ford's memoir before its official publication date. The Supreme Court

declared the action was not protected by fair use since it threatened the revenue potential of the forthcoming book thus demonstrating the significance of market effect analysis.

- Bridgeport Music, Inc. v. Dimension Films (2006): The legal dispute focused on a two-second guitar riff sample usage. The court decided against fair use after determining that any small amount of sound recording material constitutes copyright infringement. The ruling resulted in artists becoming more cautious about using music samples because copyright holders now demanded sample clearance for any amount of music used.
- The publisher made use of thumbnail images from concert posters in their Grateful Dead biography in Bill Graham Archives v. Dorling Kindersley Ltd. (2006). The court established that this qualified as fair use since the images were transformed for historical analysis rather than their original function as concert art.
- The court made a decision in Fox News Network LLC v. TVEyes Inc. (2018) about the media-monitoring service TVEyes which let users search television clips. The court declared that TVEyes' service was marginally transformative but did not qualify as fair use since it distributed complete content in a manner that threatened the original market.

The courts established through these decisions that fair use evaluations depend on the specific circumstances of each case including the reason for the use and its nature along with the extent and market implications. The judiciary continues to play a crucial role in fair use interpretation as the doctrine evolves through technological and cultural transformations. The resulting effects from these decisions spread across multiple domains to shape the way media gets produced, distributed and received by audiences. The precedential value of these cases provides direction to creators as well as educators and lawyers and courts when they apply the fair use doctrine.

6.5 The Digital Millennium Copyright Act

The Digital Millennium Copyright Act (DMCA) of 1998 is a crucial legislation in the United States that amplifies the copyright law to address the issue of managing intellectual property rights in the digital environment. The DMCA is composed of five titles, each with particular provisions:

- Title I: WIPO Copyright Treaties Implementation Title I incorporates two 1996 World Intellectual Property Organization treaties. It provides an update on the copyright law concerning the copying, distribution and modification of works in digital format and prohibits the bypassing of technological measures (anti-circumvention) that have been deployed to protect digital works. It is prohibited to produce, distribute or offer devices or services which are designed to bypass these measures.
- Title II: Online Copyright Infringement Liability Limitation Act This provision offers an immunity to online service providers (OSPs) and sets out conditions that should be met to prevent liability for copyright infringement. OSPs are required to remove or disable access to infringing material as soon as possible (notice-and-take-down procedure) and meet other requirements in order to obtain this protection.
- Title III: Computer Maintenance Competition Assurance Act Title III permits making a copy of a computer program in order to activate a computer for the purposes of maintenance or repair.
- Title IV: Miscellaneous Provisions This title provides for several miscellaneous provisions with regard to copyright management information, ephemeral recordings, distance education and libraries. It allows for the lawful distribution of encrypted programs broadcast over digital television and sets limitations on copyright for nonprofit libraries, archives and educational institutions so as to allow them to make copies for preservation purposes under specified circumstances.

- Title V: Vessel Hull Design Protection Act It's somewhat off the wall and not really related to digital copyright, but was tacked on to the DMCA. It provides protection to original designs of vessel hulls by establishing a new form of protection under copyright law.

The DMCA has had quite a number of implications for the following stakeholders:

1. Content creators and rights holders, who have tools to control the use of their digital works and a legal framework to address infringement.
2. Technology developers and manufacturers who need to consider the anti-circumvention provisions when producing devices and software that will come into contact with protected content.
3. Online service providers who enjoy the benefit of safe harbor but also have to implement adequate notice-and-take-down procedures.
4. Consumers and users, who may find that their use of media is limited by anti-circumvention provisions but also may be protected by fair use provisions that the DMCA does not abrogate.

The DMCA has been both praised and criticized. Supporters of the Act have argued that it is vital for protecting copyright in the digital age, while its opponents have argued that it can be employed to restrain competition and innovation, limit fair use and restrict freedom of speech. The anti-circumvention provisions have been particularly controversial, with arguments made about their impact on research, security testing and the right to modify or repair the products that have been purchased.

Since its implementation, the DMCA has been tested in courts, and its interpretation continues to shift. It remains a vital part of U.S. copyright law, as it continues to attempt to balance the rights of copyright owners with those of the public in the growing digital world of media use and production.

6.6 The Impact of Open-Source Software and the Sharing Economy on Intellectual Property Rights

The digital revolution brought open-source software and the sharing economy to the world which transformed information and resource creation distribution and accessibility. These developments substantially transformed the fundamental understanding of intellectual property rights. This session evaluates the consequences of open-source software and the sharing economy on intellectual property rights while discussing the current landscape's benefits and challenges and developing perspectives.

1. Open-Source Software and Collaboration: Open-source software is a system of programs that allows users to examine and modify their source code while sharing the software freely among users. The collaborative software development method produces various effects on intellectual property rights which include:
   - **(a)** Increased Innovation: Open-source software enables decentralized collaborative development models that lead to accelerated innovation and expanded software choices through global developer involvement.
   - **(b)** The sharing of open-source software enables developers to gain knowledge from existing codebases because they can review and expand upon them. The unrestricted information exchange enables knowledge acquisition and enables people to join the software development process.
   - **(c)** Open-source software receives active support from communities who deliver feedback alongside enhancements. The active communities sustain software development through ongoing maintenance which leads to its continuous enhancement.

2. The open-source software model creates both advantages for innovation and multiple obstacles which make intellectual property rights management complex.
   (a) License Compliance: Users of open-source software need to follow specific licenses (e.g., GNU General Public License) that require particular duties as well as restrictions. Open-source software users must follow license regulations by maintaining correct modification sharing procedures alongside proper authorship recognition.
   (b) The decentralized nature of open-source software creates fragmentation and compatibility challenges. The different development directions alongside incompatible version releases generate obstacles that make software integration and interoperability difficult to achieve.
   (c) Open-source software projects may unknowingly infringe on current intellectual property rights. These conflicts can result in legal disputes, potentially hindering innovation and collaboration.
3. The Sharing Economy and Intellectual Property: Several peer-to-peer platforms within the sharing economy enable users to exchange resources alongside services and expert knowledge. The new sharing economy model has disrupted multiple industries but simultaneously affects the way intellectual property rights function.
   (a) Copyright Implications: The sharing economy platforms handle the distribution of copyrighted materials including music, videos and written content to consumers. The proper sharing of resources must harmonize with copyright protection standards to maintain integrity.
   (b) Trademark and Branding: Sharing economy platforms encounter trademark-related problems in their operations. Brand name misuse without permission leads to consumer misunderstanding and devalues trademark quality.
   (c) Collaborative Consumption: The sharing economy promotes resource optimization by allowing users to share and access resources instead of buying them independently. Traditional models based on exclusivity and control encounter obstacles because of this change.
4. Evolving Perspectives and Adaptations: Open-source software and the sharing economy have led to modifications in intellectual property rights perspectives during their ongoing development.
   (a) Hybrid Approaches: Companies implement mixed development approaches which combine proprietary elements with open-source capabilities to achieve commercial objectives and collaborative innovation. This method lets organizations maintain unique differentiators through open-source collaboration benefits.
   (b) Licensing Innovations: Creative Commons alongside other licensing models emerged to empower content creators specify usage permissions and promote collaborative work. Licensing models help content creators determine sharing permissions to encourage teamwork.
   (c) Legal Frameworks: Intellectual property laws receive ongoing modifications that address issues stemming from open-source software and sharing economy operations. Courts together with legislative bodies create new legal frameworks which adapt to the developing practices in these areas.

In conclusion, the multiple effects open-source software and the sharing economy have on intellectual property rights present a complex situation. The benefits of innovation and resource optimization through these models are offset by difficulties managing licenses and protecting intellectual properties. Maintaining equilibrium between collaborative approaches and resource protection and usage remains vital. A dynamic and inclusive intellectual property ecosystem requires ongoing pursuit of legal frameworks alongside industry standards and collaborative efforts to protect creators' rights and interests and those of innovators and users.

6.7 The Role of Government in Protecting Intellectual Property

Intellectual property (IP) serves as a vital element for encouraging innovation as well as creativity and economic expansion. Governments throughout the world have created legal systems together with protection institutions to safeguard the rights and interests of creators and brand owners and inventors. The presentation demonstrates government involvement in intellectual property protection through discussion of its essential duties and innovative initiatives that support creative and innovative environments.

1. The establishment and enforcement of intellectual property rights laws depend heavily on government intervention. Key legislative frameworks include:
    1. Through copyright laws governments protect original works of authorship by implementing legal protections for literary and artistic creations as well as musical and audiovisual works. The copyright framework gives creators legal control over their work through exclusive rights to reproduce, distribute and perform their creations but includes provisions for public use exceptions.
    2. The exclusive rights of inventors to their inventions receive protection through patent laws which prevent others from making or using or selling patented inventions without authorization. The establishment of patent systems by governments serves two purposes by offering legal safeguards to innovations while requiring inventors to disclose their technological advancements.
    3. Trademark laws protect distinctive signs and logos together with symbols which help distinguish products and services from others in the marketplace. The establishment of trademark systems by governments allows them to stop consumer confusion while defending brand identities by granting exclusive rights for trademark registration and protection.
2. Government agencies responsible for enforcing intellectual property rights rely on enforcement mechanisms as their fundamental tool. These mechanisms include:
    1. The registration process of intellectual property rights such as patents copyrights and trademarks falls under specialized offices or agencies which government agencies establish for this purpose. These offices maintain legal records and act as essential guardians for the integrity of intellectual property systems.
    2. IP rights holders can use legal frameworks provided by governments to protect their rights through civil court actions. The judicial system resolves disputes through injunctions and damage awards and other legal measures that defend intellectual property rights.
    3. Governments use criminal prosecution to prevent intellectual property infringement through the threat of prosecution. Law enforcement agencies investigate cases of counterfeiting and piracy and other intentional infringement to protect IP rights through prosecution efforts that aim to stop illegal activities.
3. International cooperation remains active among governments because they work together to standardize intellectual property protection throughout national borders. Key initiatives include:
    1. World Intellectual Property Organization (WIPO): Through WIPO governments participate in a United Nations specialized agency to create international IP policies and offer technical assistance to members and help achieve IP cooperation. The development of international treaties and agreements including Berne Convention and Paris Convention and TRIPS Agreement receives support from WIPO.

2. Through bilateral and multilateral agreements governments establish international partnerships that promote coordinated IP law enforcement and harmonization. These agreements enable member countries to exchange information and establish joint enforcement systems as well as protect IP rights during international trade operations.

4. The government plays an essential part in educating people about intellectual property rights while establishing societal respect for these rights. Initiatives include:

    1. Governments implement awareness programs which teach the public and business entities and creative professionals about intellectual property significance and infringement consequences. These efforts aim to promote compliance and respect for IP rights.

    2. Governments direct their resources toward building capacity programs which improve the understanding and enforcement of intellectual property laws. The effective implementation of legal frameworks benefits from training programs that focus on judges and prosecutors as well as law enforcement officers and IP professionals.

5. Governments function as a balancing force which manages competing stakeholder interests among IP rights holders and consumers and public welfare beneficiaries. This involves:

    1. The development of government policies incorporates the multiple requirements and positions of stake-holders through their policy creation processes. The balance between creators' rights and public interests can be achieved through copyright law provisions which establish fair use exceptions.

    2. Public consultations involving industry associations along with consumer organizations and civil society groups form part of the government's stakeholder engagement approach. Government consultations with various stakeholders produce policies and regulations which strive to balance intellectual property protection through diverse stakeholder views.

In conclusion, government plays an essential role in protecting intellectual property because it supports innovation and creativity and drives economic development. Governments implement legislative frameworks together with enforcement mechanisms and international cooperation and education and stakeholder engagement to protect IP rights while ensuring knowledge and cultural expression access. Through their supportive legal framework governments support industry growth while encouraging innovation to drive economic advancement in the global knowledge-based economy.

6.8 Case Study 1

RESEARCH ARTICLE

Intuitive intellectual property law: A nationally-representative test of the plagiarism fallacy

Anne A. Fast[1]*, Kristina R. Olson[1], Gregory N. Mandel[2]

1 Department of Psychology, University of Washington, Seattle, Washington, United States of America,

2 Beasley School of Law, Temple University, Philadelphia, Pennsylvania, United States of America

* *afast@uw.edu*

Abstract

Studies with convenience samples have suggested that the lay public's conception of intellectual property laws, including how the laws should regulate and why they should exist, are largely incommensurate with the actual intended purpose of intellectual property laws and their history in the United States. In this paper, we test whether these findings generalize to a more diverse and representative sample. The major findings from past work were replicated in the current study. When presented with several potential reasons for IP protection, the lay public endorsed plagiarism and felt that acknowledging the original source of a creative work should make copying that work permissible—viewpoints strongly divergent from lawmakers' intent and the law itself. In addition, we replicate the finding that lay people know remarkably little about intellectual property laws more generally and report little experience as users or creators of creative works.

Introduction

As of January 2017, over six million videos on YouTube contain the curious message: "no copyright intended" or "no copyright infringement intended." Nearly as curious is that these videos routinely contain copyrighted material that the video poster does not have permission to use. Thus it appears that the video posters seem to believe that citing the original source of the copyrighted material, in combination with a statement that this poster is not the copyright holder, somehow reduces or eliminates the video poster's culpability for copyright infringement. However, in the United States and most countries, this behavior is often illegal [1]. It appears that hundreds of thousands of YouTube posters have a different viewpoint about the most basic tenets of copyright law. In this paper, we explore the possibility that a large swath of the American public holds the beliefs that intellectual property protection is designed to prevent plagiarism and that acknowledging the source of creative work makes it legal for a subsequent user of that work to duplicate the material without permission.

Past work has used the term *plagiarism fallacy* to refer to a set of beliefs held by American adults that suggest they have major concerns with plagiarism. That is, they believe plagiarism concerns underlie U.S. IP protection and that remediation of IP infringement is possible through, amongst other things, attribution of credit [1]. These beliefs are considered to be a *fallacy* because lawmakers and legal experts do not consider plagiarism concerns to be a basis for intellectual property protection and IP laws generally do not permit use of others' creative works simply by providing attribution [2–4]. We explore this plagiarism fallacy in the current work.

https://doi.org/10.1371/journal.pone.0184315 Used with CC 4.0 license

The legal basis for IP protection

Intellectual property law in the United States is commonly understood by policymakers and legal experts to be designed primarily to incentivize the creation of novel, transformative products of the mind [2]. According to this logic, without potential protection, creators would not invest their time in the invention and development of new medicines, artistic work, or advancing technology. Of course, the justification for IP law is not singular, and alternate views have been advocated over time. For example, some commentators have focused on John Locke's natural rights theory to argue that creators should hold *natural rights* in their intellectual creations or that by virtue of birthing a creation, the creator deserves the right to control the copying and distribution of their efforts [3]. The work of Kant and Hegel has been the basis for yet another argument, an *expressive rights* argument—the belief that providing intellectual property protection encourages cultural development and the expression of personal freedom, personality or beliefs [4]. Critically, however, recent empirical legal work suggests that the lay public may not agree with any of these arguments as the foundation for IP. Instead, the theory that appears most commonly to underlie lay views of IP protection is anti-plagiarism.

Lay intuitions on plagiarism and attribution

In one study, Mandel, Fast, and Olson (see Study 2 [1]) presented 116 adults recruited through Amazon mechanical Turk with four possible justifications for IP protection—the three views widely identified by legal experts and scholars to actually underlie IP protection in the U.S. (incentives, natural rights, and expressive rights)—along with an alternative justification—to prevent plagiarism. Inconsistent with the actual law and viewpoints of IP law experts, but consistent with the *plagiarism fallacy*, plagiarism was the most widely endorsed reason for IP protection.

Likely because people view plagiarism as the primary legal concern in the re-use of creative works, they appear to believe that attribution to the original source of material reduces culpability for its re-use. Mandel and colleagues [1] first demonstrated this effect. They presented 443 mTurk participants with a series of vignettes in which one person (User) re-used, without permission, material first created by someone else (Creator). Participants were asked how permissible this re-use was. Then participants were told to assume that the User (still without permission) had credited the Creator as the original source and were asked how permissible the re-use was in this case. In every scenario—spanning material covered by copyright laws such as music and painting to material covered by patent laws such as vaccines and software—participants believed it *was* more permissible and *should be* more permissible to re-use the material when the original source was cited than when the original source was not cited. Further, in nearly every case, the majority of participants believed that attribution made it permissible for the User to make the copy if the User simply attributed the original creation to the Creator. The ameliorative effect of attribution was replicated in two additional studies using mTurk participants [5]. Consistent with these results, young children have been shown to believe that attribution to a source makes copying more permissible [6]. Yet, importantly, in the U.S., attribution does not make re-use any more legal. In fact, all of the scenarios presented to participants in the work of Mandel, Fast, and Olson [1] would constitute intellectual property infringement under U.S. IP law. Thus, past work suggests a very clear disconnect between the actual U.S. copyright and patent law and lay people's intuitions about these laws.

Limitations of past work

While early findings demonstrated the plagiarism fallacy, there are limitations to the past work. Chief amongst these limitations is the lack of representativeness of the samples utilized. All of these prior data were collected using Amazon mechanical Turk, an online marketplace of workers who complete surveys (and other tasks) in exchange for pay. While mTurk has proven to be a useful testing ground for behavioral research [7], the lack of diversity within this sample could be a major problem. As Paolacci and Chandler [8] have warned, findings from Amazon mechanical turk "should not be treated as representative of the general population" (p. 185).

https://doi.org/10.1371/journal.pone.0184315

The specific mTurk samples used in all of the previous plagiarism work have been disproportionately young, liberal, educated, and White, relative to the national population [1,5]. These demographic characteristics could explain all or part of the effects observed. For example, past research has found that liberals and people with greater education are less concerned about money [9,10], which could lead to a lower belief in financial incentives as motivation for creative work. In addition, greater education also means higher exposure to academic norms, which tend to utilize a plagiarism-based model in which people gain reputation via citations or attribution to their work. In this way, greater exposure to academic norms may in fact increase plagiarism concerns, drowning out other concerns. Alternatively, greater education could result is greater knowledge of the law (insofar as education is on topics related to law), which runs contrary to both the plagiarism fallacy and the attribution fallacy, thus actually decreasing the biases displayed in past work. While past work did not specifically observe a correlation between education and the plagiarism fallacy or attribution fallacy, perhaps the lack of a large enough sample of conservatives and participants with less education made it difficult to observe such an effect. Investigating the same phenomena within a more diverse and representative sample is crucial if one wants to draw conclusions about Americans overall, rather than just highly educated or liberal, young Americans. The current work aims to fill this gap in knowledge.

Current work and hypotheses

This study was designed to ask whether past findings concerning the plagiarism fallacy replicate in a broader, representative sample of U.S. adults. Specifically we wanted to re-test the hypotheses that the lay public (1) believes that anti-plagiarism is the most important reason for IP law, and (2) believes that copying without permission should be more permissible if attribution to the original source of the creative work is made.

Further, for comparison to past work, we explored whether demographic differences, experience with IP, and/or knowledge about IP are associated with beliefs about plagiarism, attribution, and general views about IP protection. Previous work by Mandel and colleagues [1] as well as work by Fast and colleagues [5] found that knowledge of IP law is low and that most people report fairly limited experience with intellectual property (as either creator or user). In both papers individuals who knew more about the law were less likely to believe attribution should (or does) influence culpability when creative work is re-used by another person. Further, Mandel and colleagues [1] found that older, female, and more conservative individuals endorsed greater need for IP law compliance. Fast and colleagues [5] similarly found that older, female, and employed participants more strongly support IP protection in one study, and that older and female participants more strongly supported IP protection in a second study. We re-examine these relationships in a broader, more representative sample here.

Method

Participants

This research was approved by both the University of Washington's Human Subjects Division and Temple University's IRB. The sample included 506 adults (M_{age} = 43.67 years, SD = 14.70 years; 51% female) residing in the U.S. Participants were recruited in May 2016 through an online participant recruitment company that compensated participants for completing an online survey. The company was hired to recruit a sample that was demographically representative of the U.S. population as a whole. See Table 6.1 for detailed participant demographic information. The sample was selected to be representative of the national U.S. adult population across a variety of demographic characteristics, including gender, age, and race. Participation was completed online, thus participants were given informed consent via an information statement at the beginning of the online survey. Participants were required to agree to the information before proceeding with the survey.

https://doi.org/10.1371/journal.pone.0184315

Table 6.1 Demographics of participants.

Variable	Category	Percent of Sample
Age		
	18–29	17.7%
	30–39	33.8%
	40–59	30.2%
	60 and older	11.1%
Gender		
	Male	48%
	Female	51%
	Other	.2%
	Do not wish to report	.2%
Race		
	Non-Hispanic White or Euro-American	59%
	Black, Afro-Caribbean, or African American	19%
	Latino or Hispanic	10%
	Native American or Alaskan Native	2%
	South Asian or Indian American	5%
	Middle Eastern or Arab American	.2%
	Other or more than one	4%
Education		
	Less Than High School	1%
	Some High School	2%
	High School Grad	19%
	Some College	22%
	Trade/Technical Training	5%
	Associate Degree	12%
	Bachelor's Degree	27%
	Some Postgrad Work	2%
	Postgrad Degree	10%
Residence		
	Urban	26%
	Suburban	44%
	Small Town	14%
	Rural	16%
Political Ideology		
	Very Conservative	10%
	Conservative	19%
	Moderate	38%
	Liberal	17%

(*Continued*)

https://doi.org/10.1371/journal.pone.0184315

Table 6.1 Demographics of participants. (*Continued*)

Variable	Category	Percent of Sample
	Very Liberal	9%
	Other	.2%
	Don't know	6%
Employment		
	Unemployed	14%
	Part-Time	14%
	Full-Time	50%
	Student	4%
	Retired	13%
	Other	5%
Income		
	Less than 10,000	7%
	10,000–19,000	7%
	20,000–29,000	12%
	30,000–39,000	12%
	40,000–49,000	10%
	50,000–74,000	21%
	75,000–99,000	14%
	100,000–150,000	8%
	More than 150,000	5%
	Do not wish to report	4%

https://doi.org/10.1371/journal.pone.0184315.t001

Materials and procedure

The survey had five phases: (1) an item ranking the participant's preferred basis for IP law; (2) vignette evaluations; (3) general IP opinion items; (4) IP knowledge and experience questions; and (5) demographic information questions. Whether the ranking item came before or after the vignettes was counterbalanced but all other items were presented in the order listed above. The specific set of items used in this survey was not validated on a separate group of participants; however, all items were directly taken or adapted from previous work [1,5]. The online survey is in S1 Appendix.

Ranking potential bases of IP law. To assess the plagiarism fallacy, participants were presented with four potential reasons why there are laws regulating the products of creativity and innovation and asked to rank these four bases for IP law, ranging from what they thought is the most important (1) to what they thought is the least important (4). Each of the four justifications for IP law (incentives, natural rights, expressive rights, and anti-plagiarism) was described in one paragraph (see S1 Appendix for exact wording of this ranking item). This item was adapted from Mandel, Fast, & Olson [1].

Vignette evaluations. To assess the attribution fallacy, participants were presented with 8 vignettes reflecting a 2 (Intellectual Property Type: "expression" [copying some expression from a creative product into another work] and "product" [copying the complete creative product]) X 4 (Subject Matter: engineering, medical, music, and painting) design. The full vignettes are listed in the S1 Appendix. In each vignette, a person copied

all or part of a product that was created by another person. Following each vignette, participants were asked "In your opinion, should [the property user's] action be allowed?" Importantly, participants were asked to indicate their "personal opinion[s] about whether the action should or should not be allowed, regardless of what the law might actually be."

Participants were then asked if the action should be permissible if the user had credited the original source of the property: "In your opinion, if [property user] gives [creator/property owner] credit as the source, should [property user's] action be allowed?" Participants reported their answers to each question using a scale ranging from 1 (definitely not allowed) to 6 (definitely allowed). These questions were taken from previous work [1,5].

General IP opinion items. In order to measure participants' level of support for intellectual property law more broadly, four general IP opinion questions were included in the survey (taken from [5]). The four questions were: 1) Do you think intellectual property laws in the United States should generally be made stronger, weaker, or left about where they are?; 2) How important do you believe it is for people to comply with intellectual property rights laws?; 3) How carefully do you comply with intellectual property laws?; and 4) Intellectual property laws should be most concerned with the rights of the: creator vs. user?. Participants responded to the questions using a slider scale ranging from 0 to 100, which was anchored at 50 when the item was presented to participants. For the first three questions, the lower end of the scale represented support for weaker IP protection and the higher end support for stronger IP protection (anchors: question 1, weaker…stronger; question 2, not important…important; and question 3, not carefully…carefully). For the fourth question, the lower end of the scale represented strong support for the creator of IP, whereas the higher end of the scale represented support for the consumer of IP (anchors to question 4: creator…user). This final item was reverse-scored in order for the scale to conceptually align with the other three items.

IP knowledge and experience questions. Participants were asked ten multiple choice questions about intellectual property law and two questions about their level of experience with IP creation or use (taken from [1,5], see those papers for full list of items). These items were included to assess their actual knowledge and training with regard to IP law. For example, in the knowledge quiz, participants were asked "What is permissible under copyright law, in general, concerning material found on the Internet: (a) It can be copied to other websites or downloaded freely; (b) It can be copied to other websites freely, but not downloaded; (c) It can be copied to other websites if attribution to the original site is provided; or (d) It can be copied to other websites if the author grants permission." With regards to experience, participants responded to questions about their level of experience as a creator or producer of works protected by intellectual property and a user of intellectual property on five-point scales ranging from no experience (1) to considerable experience (5).

Demographic questions. The demographic section of the survey included questions about participant age, gender, race, education, employment status, income, political ideology, and type of residence. Responses to demographic questions are presented in Table 6.1.

Results

Ranking potential bases of IP law

Participants' rankings of the possible bases for IP protection were not random, $\chi^2(3) = 60.59$, $p < .001$. Specifically, participants ranked Plagiarism (39%) as the most important reason for having IP laws more often than the other rationales (Natural rights: 24%; Incentives: 21%; Expressive rights: 16%). Expressive rights (36%) were most likely to be ranked least important compared to the other three bases (Incentives: 24%; Natural rights: 21%; Plagiarism: 19%), again with these answers differing from random chance, $\chi^2(3) = 33.06$, $p < .001$. See Table 6.2 for a full breakdown of participant rankings.

https://doi.org/10.1371/journal.pone.0184315

Table 6.2 Participant rankings of potential reasons for IP law.

	Ranked 1st	Ranked 2nd	Ranked 3rd	Ranked 4th
Plagiarism	39%	24%	18%	19%
Natural Rights	24%	32%	22%	21%
Incentives	21%	25%	30%	24%
Expressive Rights	16%	19%	30%	36%

https://doi.org/10.1371/journal.pone.0184315.t002

Vignette evaluations

Responses to the vignettes were analyzed in a 2 (Intellectual Property Type: expression vs. complete product) x 4 (Subject Matter: music vs. painting vs. engineering vs. medical) x 2 (Mitigating Factor: baseline vs. attribution) mixed measures ANOVA with Intellectual Property Type as a between-subjects factor and Subject Matter and Mitigating Factor as within-subjects factors (the S1 File provides additional information about the study design and analyses). Significant effects were followed-up with pairwise Bonferroni-corrected comparisons. The percent of participants responding that the actions should be allowed (by providing a response of 4, 5, or 6 on the scale) are presented in Table 6.3.

Table 6.3 Percentage of participants responding "allowed" to vignette subject matters.

	Baseline	Attribution
Music	49%	67%
Painting	58%	73%
Medical	48%	66%
Engineering	50%	62%

Percentage of participants who believed copying behaviors in the vignettes should be "allowed." If a participant chose a 4, 5, or 6 on the 1 (definitely not allowed) to 6 (definitely allowed) scale, they were coded as saying the behavior should be "allowed."

https://doi.org/10.1371/journal.pone.0184315.t003

Analyses revealed a significant main effect of Intellectual Property Type on evaluations, $F(1,499) = 62.90$, $p < .001$, $\eta_p^2 = .112$, such that participants believed, overall, that copying the complete creative product ($M = 3.36$, $SD = 1.13$) should be less permissible than copying some, but not all, of the expression ($M = 4.15$, $SD = 1.12$). Analyses also revealed a significant main effect of Subject Matter on evaluations, $F(3,1497) = 12.38$, $p < .001$, $\eta_p^2 = .024$. Overall, participants found copying in the painting vignette ($M = 4.02$, $SD = 1.39$) less permissible than reusing material in the other three items (music: $M = 3.66$, $SD = 1.54$; engineering: $M = 3.68$, $SD = 1.48$; medical: $M = 3.66$, $SD = 1.57$), all $ps < .001$. Responses to the other three Subject Matters (music, engineering, and medical) did not differ from one another, all $ps = 1.0$. Finally, there was a significant main effect of Mitigating Factor on evaluations, $F(1,499) = 158.94$, $p < .001$, $\eta_p^2 = .242$, consistent with the Attribution fallacy, such that using someone else's property without providing attribution to the original source ($M = 3.46$, $SD = 1.25$) was seen as less permissible than using someone else's property while providing attribution to the original source ($M = 4.05$, $SD = 1.21$), $p < .001$.

These main effects were qualified by multiple interactions. Analyses revealed a significant interaction between Intellectual Property Type and Subject Matter, $F(3,1497) = 20.05$, $p < .001$, $\eta_p^2 = .039$. Participants saw it as less permissible to copy a complete product than the expression of a product within the painting, engineering, and medical vignettes, all $ps < .001$, but not for the music vignette, $p = .256$.

https://doi.org/10.1371/journal.pone.0184315

Further, analyses revealed a significant interaction between Subject Matter and Mitigating Factor, $F(3,1497) = 3.98$, $p = .008$, $\eta_p^2 = .008$. While in all cases participants thought it was more permissible to copy the creator's work when attribution was given than when it was not, the degree of this difference between attribution and baseline differed by subject matter: music $M = .672$, engineering $M = .685$, painting $M = .517$ and medical $M = .451$.

Finally, there was a significant three-way interaction between Intellectual Property Type, Subject Matter and Mitigating Factor, $F(3,1497) = 6.79$, $p < .001$, $\eta_p^2 = .013$. Specifically, attribution increased permissibility compared to the baseline condition more in the expression condition compared to the physical product condition for music vignette (mean differences = .835 and .510, respectively); in contrast, attribution increased permissibility compared to the baseline condition more in the complete product condition compared to the expression condition for the painting (mean differences = .629 and .405, respectively), engineering (mean differences = .792 and .579, respectively) and medical (mean differences = .629 and .273, respectively) vignettes. That is, although there was an attribution fallacy for all vignettes, the degree of that effect differed depending on Domain and Property Type.

General IP opinion items

Participants' responses to all four of the general IP opinion questions were significantly higher than the midpoint of the scale (50), suggesting that participants tended to: (1) favor stronger IP laws, $t(498) = 22.65$, $p < .001$; (2) support compliance with IP laws, $t(498) = 36.30$, $p < .001$; (3) report self-compliance with IP laws, $t(496) = 30.41$, $p < .001$; and (4) support laws that are concerned with the rights of creators of IP, $t(496) = 4.97$, $p < .001$. Means and standard deviations for each of the general IP opinion items are listed in Table 6.4. Participants' scores on the first three opinion questions were highly related to one another (Cronbach's alpha = .74); therefore, we computed a factor called *IP Support*, which is the average of scores on these three items. Higher IP Support scores indicate greater support for the strength of and compliance with IP law. This IP Support factor was used to assess individual difference analyses below. The fourth item reduced the reliability of the composite and therefore was excluded from subsequent analyses.

Table 6.4 Means of general IP opinion questions.

	M (SD)
IP laws in the USA should be stronger, weaker, or the same?	69.66 (19.39)
How important is it to comply with IP laws?	81.06 (19.11)
How carefully do you comply with IP laws?	79.60 (21.70)
Laws should be concerned with rights of: creator vs. user	57.76 (34.78)

https://doi.org/10.1371/journal.pone.0184315.t004

IP knowledge and IP experience

An IP Knowledge score was calculated for each participant, representing the number of correct responses on the 10-item intellectual property quiz ($M = 3.29$, $SD = 1.57$). Participants tended to perform better than chance (2.5 out of 10 correct) on the IP knowledge quiz, $t(501) = 11.22$, $p < .001$, though performance was still poor (well below 50%), consistent with the findings reported Mandel et al [1].

Participants reported having little to no experience as a creator or producer of IP ($M = 1.80$, $SD = 1.18$) or as a user of IP ($M = 2.00$, $SD = 1.34$), again consistent with Mandel et al [1]. These items were significantly positively correlated, $r = 0.73$, $p < .001$, thus we averaged these items to create an IP Experience score for each participant.

Individual difference analyses

We examined the correlations between individual difference measures and the main dependent measures. Importantly, because of the large number of correlations conducted, which inflates Type I error rates, we assessed significance of correlations with a p-value cutoff of $p < .01$. We first paid specific attention to whether the previously reported findings [1,5] replicated. Unlike in both previous papers, we found that IP knowledge was not significantly associated with responses to the attribution items. That is, people who knew more about IP as demonstrated by our quiz were no more or less likely to believe that attribution should affect the legal permissibility of copying. In addition, while past work [1,5] found that older, female, more conservative and employed participants more strongly support IP protection, in this more diverse, representative sample, none of the demographic variables was significantly correlated with IP support.

Discussion

Overall, we found, in a large replication study, that previous findings reporting the plagiarism fallacy are robust in a more diverse and representative sample. Specifically, many Americans believe that the most important reason why IP rights exist is to prevent plagiarism, rather than to create incentives for creative work, to promote expressive rights, or to honor natural rights. Further, and likely as a result of this concern for plagiarism, adults believe that acknowledging the source of creative work should make copying that work more permissible, regardless of actual permission. In fact, the majority of participants in the majority of scenarios believed that copying others' work (without permission) should be permissible as long as the user attributed the source. These findings suggest that past work with demographically-limited samples [1,5] are not only replicable but replicable in more diverse samples. In recent years, a push to replicate major findings and to assess their generalizability has been advocated across fields of science [11–14].

In addition to the primary findings, we examined several other findings from the past literature. Two additional findings that replicated were the observation that, in general, the lay public knows very little about intellectual property law and that the lay public feels that they have limited experience as users or creators of creative products. We also re-investigated the relationship between our primary dependent variables and a range of demographic variables. None of the primary individual difference relationships replicated. Thus, we believe that the main effects—the plagiarism fallacy and the attribution fallacy—are more robust and do not appear to rely on the representativeness (or lack thereof) of the sample. The individual differences however, appeared less reliable across studies with different demographic compositions.

Although the current work had a much more diverse sample of participants compared to past work, the current study only included participants who reside in the United States. Thus one limitation of this work is that it reflects the views of Western, educated, industrialized, rich, and democratic (WEIRD) individuals, whose opinions likely do not generalize to non-WEIRD individuals [15]. Future work could explore the tendency to endorse the plagiarism fallacy across cultures. Another potential limitation of the current study is that the vignettes were directly drawn from previous studies, making it difficult to know whether participants' consistent responses to the vignettes from past to current work depend on those particular scenarios, or whether their responses would generalize to other scenarios and even other domains of infringement. Therefore, future work could investigate individuals' responses to copying in new scenarios and domains.

Another limitation in our study design was our use of the ranking item for assessing people's beliefs about the basis of IP protection. Specifically, we provided four possible reasons, but participants may have had other ideas not represented by these four possibilities. Thus, we can only conclude that plagiarism was the most common amongst these four options rather than that plagiarism would necessarily have been the most endorsed in a range of other options. We are skeptical that this is a major limitation, however, because Mandel and colleagues [1] did ask a more open-ended question and found that participants frequently indicated concerns

about plagiarism as the reason for concerns in infringement cases. Further, our findings that attribution made copying more permissible suggest that plagiarism concerns are important to participants. However, future work might address this concern with more open-ended responses.

Despite these limitations, our replication of the lay public's beliefs concerning attribution is particularly notable, as all of the scenarios in this study were specifically selected because these actions constitute intellectual property rights infringement under current U.S. IP law. While about half of respondents endorsed behavior that violates IP law on the baseline item, respondents were significantly more likely to do so on the attribution item. That is, the majority of respondents on most of the vignettes believed that providing attribution to the original source should make copying behavior permissible, even though the law and lawmakers do not permit such behavior. Such a finding is critical for legal scholars and policy-makers to understand. Where individuals disagree with a law, they are less likely to view the law as legitimate and less likely to comply with the law [16].

The attribution finding is also worthy of note in light of recent findings about creators' preferences. Sprigman, Buccafusco, and Burns [17] found that creators themselves appeared to value attribution. In their study, creators were willing to accept less pay in exchange for attribution of their work. Thus it appears that not only third party observers, with little experience in creating these works, like the participants in the current work, but also actual creators, value attribution in a way that U.S. law does not. Intriguingly, many European countries provide for certain attribution rights for authors under their copyright laws [18]. This system may in fact better align with the way American adults intuitively reason about creative works.

For legal systems to work, people governed by that system need to know, understand, and follow that governance. The current work suggests that beliefs about intellectual property protection that differ from those of legal experts extend beyond those individuals who create videos on YouTube or participate in behavioral experiments on Amazon mechanical Turk, and instead are fairly representative of the American public. Such findings should be discomforting to policy-makers and suggest the need for greater education and intervention, or changes in legal policy.

Supporting information

S1 Appendix. Online survey.
(DOCX)

S1 File. Additional study design & analysis information.
(DOCX)

Acknowledgments

The authors wish to thank Diana Joskowicz for her research assistance.

Author Contributions

Conceptualization: Anne A. Fast, Kristina R. Olson, Gregory N. Mandel.
Data curation: Anne A. Fast.
Formal analysis: Anne A. Fast.
Funding acquisition: Kristina R. Olson, Gregory N. Mandel.
Investigation: Anne A. Fast.
Methodology: Anne A. Fast, Kristina R. Olson, Gregory N. Mandel.

Project administration: Kristina R. Olson, Gregory N. Mandel.
Resources: Kristina R. Olson, Gregory N. Mandel.
Writing – original draft: Anne A. Fast, Kristina R. Olson.
Writing – review & editing: Anne A. Fast, Kristina R. Olson, Gregory N. Mandel.

References

1. Mandel GN, Fast AA, Olson KR. Intellectual property law's plagiarism fallacy. BYU Law Rev. 2016; 915–983.
2. Burk DL, Lemley MA. Policy Levers in Patent Law, Virg Law Rev. 2003; 1575 (89): 1597–99.
3. Hughes J. The Philosophy of Intellectual Property, Georgetown Law Jour. 1988; 287 (77): 296–330.
4. Radin MJ. Market-inalienability. Harv. L. Rev. 1987; 1849 (100): 1894–95.
5. Fast AA, Olson KR, Mandel GN. Experimental investigations on the basis for intellectual property rights. Law Hum Beh. 2016; 40(4): 458–476.
6. Shaw A, Olson KR. Whose idea is it anyway?: The importance of reputation in acknowledgement. Develop Sci. 2015; 18: 502–509.
7. Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality data? Persp. Psych Sci. 2011; 6(1): 3–5.
8. Paolacci G, Chandler J. Inside the Turk: Understanding Mechanical Turk as a participant pool. Curr Dir Psych Sci. 2014; 23(3): 184–188.
9. Furnham A. Many sides of the coin: The psychology of money usage. Pers Indiv Diff. 1984; 5(5): 501–509.
10. Furnham A. Why do people save? Attitudes to, and habits of saving money in Britain. Jour. Appl. Soc. Psych. 1985, 15(5): 354–373.
11. Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA. An open investigation of the reproducibility of cancer biology research. Elife. 2014, 3: e04333.
12. Huffmeier J, Mazei J, Schultze T. Reconceptualizing replication as a sequence of different studies: A replication typology. Jour Exp Soc Psych. 2016, 66: 81–92.
13. Nature Editorial. Go forth and replicate! Nat. 2016, 536 (373), https://doi.org/10.1038/536373a
14. Roediger HL III. Psychology's Woes and a Partial Cure: The Value of Replication. APS Observer. 2012, 25(2): 9.
15. Henrich J, Heine SJ, Norenzayan A. Most people are not WEIRD. Nature. 2010; 466: 29. https://doi.org/10.1038/466029a PMID:20595995
16. Tyler TR. Procedural Justice, Legitimacy, and the Effective Rule of Law. Crime Just. 2003, 30: 283–357
17. Buccafusco CJ, Sprigman CJ, Burns KC. What's a name worth?: Experimental tests of the value of attribution in intellectual property. Bos Univ Law Rev. 2012, 93: 1.
18. Ginsburg JC. European Copyright Code—Back to First Principles (With Some Additional Detail). Copyright Soc USA. 2010, 58: 265.

6.9 Case Study 2

COMMUNITY PAGE

The Open Source Seed Licence: A novel approach to safeguarding access to plant germplasm

Johannes Kotschi [1*], **Bernd Horneburg** [2*]

1 Agrecol, Guggenhausen, Germany, 2 Division of Plant Breeding, CiBreed, University of GÖttingen, GÖttingen, Germany

* *kotschi@agrecol.de (JK); bernd.horneburg@uni-goettingen.de (BH)*

Abstract

The laws to secure intellectual property rights on plant germplasm have been strongly developed in parallel to the ongoing seed market consolidation. Germplasm as a commons, i.e., a natural resource accessible to all members of a society, receives almost no legal protection. On the other hand, the use of germplasm and released cultivars in breeding is increasingly restricted by intellectual property rights. In this study, approaches to open source plant germplasm are discussed, and the Open Source Seed (OSS) Licence is introduced and analysed. The OSS Licence was developed by an interdisciplinary working group of plant breeders, agricultural scientists, lawyers, and commons experts in Europe. The aim is to protect germplasm as a commons, support the free exchange of germplasm, stimulate plant breeding, reduce costs, and accelerate innovation. The OSS Licence is a legal tool and novel approach that extends its reach on derivatives of licenced germplasm. It is compatible with current seed laws. Effects on the access to plant germplasm, on breeding for diverse pedoclimatic environments, socioeconomic systems, and on biodiversity as a whole can first show after a few breeding cycles. The impact of open source germplasm on these aspects needs to be monitored carefully.

Introduction

For millennia, crop seed has been a common good [1], and all over the world, crops have been cultivated, enhanced, and bred by farmers, a practice that resulted in a rich inter- and intraspecific diversity of crops. The emergence of agricultural sciences at the end of the 19th century marks an historical change to separating plant breeding from plant production [2]. Simultaneously, scientific plant breeding began contributing decisively to agricultural intensification. In the 1940s, combining high-yielding crops with extensive use of mineral fertilizers and chemical plant protection became the main strategy for food security [3,4]. High-yielding cultivars were developed using breeding techniques such as hybrid breeding, facilitated through cytoplasmic male sterility systems that prevent the use of farm-saved seed [5]. Uniform cropping systems with only a few crops and cultivars spread over large areas. At later stages, intellectual property rights such as plant variety protection and utility patents were implemented [6,7]. Increasing market consolidation led to a few transnational corporations dominating the global seed market [8,9]. Consequently, the number of breeding programs was greatly reduced, and decreasing competition between companies has restricted innovation [10,11,12]. There is also a significant loss of cultivars caused by consolidation, as companies reduce portfolio size after mergers. Economies of scale act in favor of large-scale propagation of single cultivars and against breeding for diverse pedoclimatic conditions. The access to plant genetic material is increasingly constrained for farmers, plant breeders in public institutions [13], small breeding companies, and public domain initiatives.

https://doi.org/10.1371/journal.pone.0184315 Used with CC 4.0 license

Thus, after genetic bottlenecks significantly reduced diversity from early domesticates to modern cultivars, [14] the next bottleneck has been created by intellectual property rights and market consolidation.

The loss of agrobiodiversity is critical. Plant genetic diversity is essential for food security as a means of adaptation to variable weather conditions, as it enhances social and ecological resilience against extreme events [15,16,17]. A rich diversity of crops and cultivars is required for the world's innumerable agroecological sites and corresponding farming systems. Cultivars are needed that not only satisfy the needs of large-scale farming but also give smallholders—by far the majority of farmers in the world—a livelihood that allows them to contribute adequately to global food supply in quantity and quality [3,18]. The demand cannot be met by the private seed sector alone [11].

While public sector plant breeding has undergone a steady decline in Europe [19], new stakeholders have evolved. During the past 40 years, public domain initiatives, mainly in the context of organic agriculture, have been established [20]. In many countries of the Global South, participatory plant breeding with smallholders has been established successfully [21,22,23]. Also, in large international institutions of the Consultative Group on International Agricultural Research (CGIAR), open access systems have been effectively applied for crop improvement [24,25]. Most of the organic plant-breeding programs in Europe are organised by associations, foundations or informal networks and operate on a nonprofit basis. Most new cultivars are released "open access," i.e., without intellectual property rights, unconditionally and to everybody. This allows users to privatize subsequent derivatives or reproductions of that plant germplasm. In other words, commons are created but not protected.

Although the International Treaty on Plant Genetic Resources for Food and Agriculture [26] recognizes so-called "farmers' rights" to use farm-saved seed, much tension has developed between farmers' rights and intellectual property rights of the private sector, in particular plant variety protection and patenting, in international governance [27,28]. Access to plant germplasm for the nonprivate seed sector is not sufficiently granted.

As a new tool, the Open Source Seed Licence for plant germplasm (new cultivars, breeding lines, and populations) was developed by an interdisciplinary working group of the nongovernmental organization (NGO) Association for AgriCulture and Ecology (Agrecol). The primary target area is Europe. We will describe and discuss i) approaches to create open source plant germplasm, ii) the Open Source Seed Licence and the steps needed for its implementation, iii) the analysis of existing seed laws for potential conflicts with open source plant germplasm, and iv) potential effects of the Open Source Seed Licence on breeding for diverse pedoclimatic and social environments and on access to plant germplasm.

Approaches to create open source plant germplasm

The term "open source" is derived from computer science, for which it refers to free access to computer source codes. The open source concept is a development of the free software principles first articulated in the GNU General Public Licence [29] by Richard Stallman, already declared in the GNU Manifesto [30]. The aim was to allow developers and users of computers to investigate how software works and be able to modify and share their work freely with others. Several authors have suggested the analogy of plant breeding with software development [31–34]. The open source rules for plant germplasm imply that anybody may use plant germplasm for production, breeding, multiplication, and distribution. At the same time, any user is obliged to impose the same rights on future owners of the plant germplasm. Any limitation beyond this is prohibited. This clause is also called "copyleft." Thus, open source is clearly distinguished from open access, which implies unlimited access. Essentially, open source safeguards access to a common good and is therefore creating a protected commons preventing privatization [35]. During the past six years, approaches to create

open source plant germplasm have evolved in the United States [36], in Germany [37], in India [38], and in East Africa [39].

The Open Source Seed Initiative (OSSI) was founded 2012 in the US by a group of academic and nonacademic plant breeders, farmers, nonprofit agencies, seed advocates, and policymakers as the first organization for the development of open source seeds [36]. Inspired by Michaels [31], who had already presented the idea of a General Public Licence for plant germplasm in 1999, OSSI developed a licence. But the licence proved too cumbersome for practical use, and subsequently, the OSSI Pledge was developed as a social mechanism that requests an ethical commitment from seed users to comply with the open source rules [13]. The pledge is used by dozens of plant breeders and seed companies internationally to create an "ethical sharing and distribution mechanism for plant breeding" [36].

The Dutch NGO Hivos started the "Open Source Seed Systems Programme" [39]. It involves cooperation with partners in East Africa and India. It is Hivos' strategy to work through multistakeholder initiatives to search for alternatives, build viable business models for open source seeds, and seek support through lobbying and advocacy. As yet, no legal and/or ethical commitment is being used.

The Apna Beej Network in India also practices an "Open Source Seed System." It works with a "material and knowledge transfer agreement" that allows members of the network to use their plant germplasm freely within the network [38].

The German NGO Agrecol started work on open source plant germplasm systems in 2012 [37]. An interdisciplinary working group of agricultural scientists, plant breeders, and lawyers compared ethical approaches using a pledge with legal approaches using a licence. A pledge seemed to meet acceptance by potential users more easily than a licence. On the other hand, it is less binding than a licence, as it might be withdrawn or disregarded. In addition, a licence meets favorable conditions for legal enforcement in the European Union (EU): according to the Nagoya Protocol and unlike in the US, the mandatory documentation of germplasm that is used for new cultivars facilitates tracing licence infringements. Consequently, the Open Source Seed (OSS) Licence was developed and published in 2016 [40]. This in turn inspired apiculturists from Apimondia, the World Beekeepers Association, to use an open source licence for honey bees [41], and the first animal genetic open source licence was developed.

The Open Source Seed Licence

The OSS Licence grants users many freedoms and demands little: the licensee is obliged to grant the same rights he or she has enjoyed to future owners of the plant germplasm. This obligation is "viral" and includes all derivatives of the germplasm. With the first licensing, a chain of contracts is started, which, in principle, is endless. Licensees become licensors, who pass on the germplasm with the same licence.

Legally, the OSS Licence is a "sui generis contract" and falls under the General Business Terms and Conditions of German Civil Law [42], a prewritten contract for general, multiple, and unilateral use by a single party and not individually negotiated. The licence is a material transfer agreement for seeds and vegetative parts of plants. It confers user rights together with the material object and its genetic information. When the material is transferred, a contract is established that ensures the rights and duties associated with the material in question as well as all future derivatives and reproductions of that plant germplasm. The OSSI pledge also intends to do this, but unlike a pledge, the licence can be used to legally protect newly developed cultivars, breeding lines, populations, and other germplasm. Thus, the open source character of germplasm can be legally enforced. This is a novelty.

https://doi.org/10.1371/journal.pone.0184315

To transfer germplasm under the OSS Licence, the licence conditions must, unambiguously, be disclosed. This means that any transfer is valid if the licensee is able to become aware of the terms and conditions of the licence. For traders, this means that an abridged version of the OSS Licence must be printed on the wrapping of small bags with reference to the online full-text version. For larger quantities, the licensors must ensure that a copy of the licence accompanies the material being transferred, and they must explicitly inform the recipient (the licensee). Licence infringements can be examined through a combination of measures. The most important one is to analyze the breeding process that the new cultivar has undergone. According to the Nagoya Protocol [43], a detailed documentation for each breeding process is mandatory. The user of plant genetic resources must record the time and place of access and, where appropriate, also examine "the presence or absence of rights and obligations relating to access and benefit sharing" [44]. Thus, the Nagoya Protocol is a strong lever to enforce the OSS Licence. Additionally, phenotypic and genotypic data can be used. Monitoring of licence infringements is considered the responsibility of all users to ensure that open source remains open source. OpenSourceSeeds aims to initiate a social process to this effect because licence infringements could have negative repercussions for everyone.

The OSS Licence was implemented on April 26, 2017, when Agrecol launched the service provider OpenSourceSeeds [45] and presented the first OSS-licensed cultivars in Berlin, Germany. OpenSourceSeeds, under the legal entity of Agrecol, licenses newly developed cultivars on behalf of breeders. It documents OSS licensed germplasm in a database and stores samples of plant germplasm. In addition, the development of value chains with OSS-licensed cultivars is supported.

OpenSourceSeeds started with two cultivars, a vegetable and a field crop. The tomato Sunviva was bred by the network of the Organic Outdoor Tomato Project launched in 2003 by the University of Göttingen. The project is based on participation and the free exchange of tomato genotypes and related knowledge [46]. The project aims to improve plant-breeding methodology [47]. Culinaris, a seed company that exclusively sells seeds free of intellectual property rights, is in charge of legal issues and maintenance breeding. The spring wheat Convento C was developed by Forschung & Züchtung Dottenfelder Hof in Bad Vilbel, Germany. Meanwhile, seven cultivars of different crops including maize have been licensed.

Discussion and outlook

The laws on securing intellectual property rights on plant germplasm have been strongly developed, whereas plant germplasm as a commons receives almost no legal protection. Thus, the possibility to license new cultivars as open source is new and unfamiliar. The release of the first OSS-licensed cultivars will stimulate further discussions and generate valuable experiences on the feasibility, acceptance, and impact of open source germplasm among plant breeders, seed producers, growers, plant genetic resources experts, and society as a whole.

Being a legal tool, the question may arise whether the OSS Licence is in conflict with existing seed laws on seed marketing and plant variety protection. In the EU, national laws like the German Seed Marketing Act are based on EU regulations. The German Seed Marketing Act regulates the release of most agricultural and horticultural crops. The act is not in conflict with OSS licensing, but germplasm of the relevant crops can only be legally released when the cultivars are registered in the Plant Variety Database of the European Commission. The European Plant Variety Protection Act is not in conflict with open source germplasm because it only applies for cultivars under plant variety protection.

Another issue to be addressed is the funding of a future plant breeding that needs to supply cultivars adapted to very diverse pedoclimatic conditions as a service for society. Financing models are to be further developed that are not based on royalties on intellectual property rights. In many developing countries, plant breeding does not follow a business model based on royalties, and even in industrial countries, there are private breeding

https://doi.org/10.1371/journal.pone.0184315

companies that do not rely on intellectual property rights. Presently, many organic plant breeders in Europe finance their breeding work partly through "cultivar development contributions" that are negotiated between breeders, seed producers, and farmers. Some have, in cooperation with the food trade, developed a levy on food items, and most of them raise funds from government programs and foundations [48]. The funds for commons-based plant breeding are still small but continuously increasing.

Incentives for the application of the OSS Licence may be manifold. At present and after the first year of introduction, a significant demand from consumers can be observed. Interviews indicate that the public appreciates the possibility to contribute to an alternative to the dominant system of privatization and market concentration in the seed sector. Exemplarily, this can be observed with the OSS-licensed tomato cultivar Sunviva. Due to the growing demand in horticulture, more and more traders are offering Sunviva seed and its seedlings. Consumers could, therefore, create a significant pull effect. Other incentives may be identified in the following years. We expect that the increasing shortage of freely available breeding material will create a strong incentive for small and medium plant breeders to license their new cultivars open source and to contribute to a commons-based pool of plant germplasm.

Expected impacts of the licence will have to be assessed in the following areas. Open source systems aim to allow plant breeders to operate and cooperate freely [36]. It is assumed that a free exchange of germplasm stimulates plant breeding, reduces costs for breeding, and accelerates innovation. This might counteract the current trend of consolidation in the seed market and increase number and diversity of breeding programs. New plant-breeding methods like participatory plant breeding might be facilitated, which can offer cheap and fast methods of cultivar development [22]. All of these aspects can be essential to sustain diversity of plant germplasm.

We expect insights into the acceptance of the OSS Licence in the next few years because breeders and researchers can decide immediately how they release new genotypes. Monitoring the effects of licensing open source plant germplasm on other areas of major importance, however, will last much longer. Effects on the access of various stakeholders to plant germplasm, on breeding programs for diverse pedoclimatic and social environments and on biodiversity as a whole can first show after a few breeding cycles. A network meeting to evaluate first experiences with the OSS Licence will take place in 2019. We expect specific effects for different crop types (minor or major crop, field crop, vegetable, fruit trees), types of cultivars (pure line, multiline, population), and breeding systems (autogamous, allogamous, clones). The impact of open source germplasm on the areas discussed above will have to be monitored carefully.

Acknowledgments

We would like to thank our colleagues who took the first steps toward open source plant germplasm systems, and we gratefully acknowledge support by the Software AG Foundation. Many thanks to James R. Myers for discussions, to Petr Karlovsky for valuable comments, and to Paul Muto for English proofreading.

References

1. Brush SB. Farmers' Bounty. Locating Crop Diversity in the Contemporary World. Yale University Presss. 2004.
2. Kloppenburg JR. First the seed. The political economy of plant biotechnology. University of Wisconsin Press. 2004. Second Edition.
3. IAASTD. Agriculture at a Crossroads. IAASTD, International Assessment of Agricultural Knowledge, Science and Technology Development. 2009. Synthesis Report.

4. Loos J, Abson DJ, Chappell MJ, Hanspach J, Mikulcak F, Tichit M, Fischer J. Putting back meaning into "sustainable intensification."Frontiers in Ecology and the Environment 2014. https://doi.org/10.1890/130157
5. Phillips SL, Wolfe MS. Evolutionary plant breeding for low input systems. Journal of Agricultural Science. 2005. 143: 245–254.
6. UPOV. International Convention for the Protection of New Varieties of Plants of December 2, 1961, as Revised at Geneva on November 10, 1972, on October 23, 1978, and on March 19, 1991. International Union for the Protection of New Varieties of Plants. 1991. Available from: http://www.upov.int/upovlex/en/conventions/1991/content.html Cited 29 August 2018
7. WTO. Uruguay Round Agreement: TRIPS. World Trade Organization 1994. Available from: https://www.wto.org/english/docs_e/legal_e/27-trips_03_e.htm Cited 29 August 2018.
8. Heinrich Böll Foundation, Rosa Luxemburg Foundation, Friends of the Earth Europe. Agrifood Atlas. Facts and figures about the corporations that control what we eat. Heinrich Böll Foundation, Berlin. 2017.
9. Howard PH. Intellectual property and consolidation in the seed industry. Crop Science. 2015. 55: 2489–2495.
10. Heller M, Eisenberg R. Can Patents Deter Innovation? The Anticommons in Biomedical Research. Science. 1998. 280 (5364): 698–701. PMID:9563938
11. Boyle J. The Second Enclosure Movement. Law and Contemporary Problems. 2003. 66 (1): 33–74.
12. Boyle J. Enclosing the Genome: What Squabbles over Genetic Patents Could Teach Us. Advances Genetic. 2003. 50: 97–122.
13. Kloppenburg J. Re-purposing the master's tools. The open source seed initiative and the struggle for seed sovereignty. Journal of Peasant Studies. 2014. 41, 1225–1246.
14. Tanksley SD, McCrouch SR. Seed banks and molecular maps: Unlocking genetic potential from the wild. Science 1997. 277:1063–1066. PMID:9262467
15. Frison EA, Cherfas J and Hodgkin T. Agricultural biodiversity is essential for a sustainable improvement in food and nutrition security. Sustainability. 2011. 3, 238–253.
16. Narloch U, Drucker AG, Pascual U. Payments for agrobiodiversity conservation services for sustained on-farm utilization of plant and animal genetic resources. Ecological Economics. 2011. 70, 1837–1845.
17. Chappell MJ, La Valle LA. Food security and biodiversity: can we have both? An agroecological analysis. Agric Human Values. 2011. 28, 3–26.
18. De Shutter O. The right to food. Seed policies and the right to food: enhancing agrobiodiversity and encouraging innovation. United Nations General Assembly. 2009. Report of the special Rapporteur on the right to food. Note by the Secretary General. A/64/170.
19. EASAC, European Academies Science Advisory Council. Genomics and crop plant science in Europe. May 2004. EASAC policy report 02, ISBN 0 85403 6008.
20. Lammerts van Bueren ET, Messmer MM, Spiele AK, Wilbois KP, editors. Organic Plant Breeding: What makes the difference? FIBL. 2011. Proceedings of the ECO-PB 10 years Anniversary Conference.

21. Murphy K, Lammer D, Lyon S, Carter B, Jones SS. Breeding for organic and low-input farming systems: An evolutionary-participatory breeding method for inbred cereal grains. Renewable Agriculture and Food Systems. 2005. 20:48–55.

22. Ceccarelli S. Efficiency of plant breeding. Crop Science. 2014. 55:1–11.

23. Bachmann L, E Cruzada, S Wright. Food Security and Farmer Empowerment. A Study of the Impacts of Farmer-Led Sustainable Agriculture in the Philippines, Laguna, Philippines. MASIPAG. 2009.

24. Byerlee D, Dubin HJ. Crop improvement in the CGIAR as a global success story of open access and international collaboration. International Journal of the Commons. 2010. 4(1): 452–480.

25. Dedeurwaerdere T. Institutionalizing global genetic resource commons for food and agriculture. In: Halewood M, López Noriega I, Louafi S, editors. Crop Genetic Resources as a Global Commons. Challenges in International Law and Governance. Routledge. 2013. pp. 368–391.

26. FAO. International Treaty on Plant Genetic Resources for Food and Agriculture. Food and Agriculture Organization of the United Nations. 2009. Available from: http://www.fao.org/plant-treaty/overview/texts-treaty/en/. Cited 29 August 2018.

27. Christinck A, Tvedt MV. The UPOV Convention, Farmers' Rights and Human Rights. An integrated assessment of potentially conflicting legal frameworks. 2015. Available from https://www.giz.de/fachexpertise/downloads/giz2015-en-upov-convention.pdf Cited 29 August 2018.

28. Andersen R. Protecting Farmers' Rights in the global IPR Regime. Policy Brief No. 15. 2007, South Asia Watch on Trade, Economics & Environment (SAWTEE).

29. FSF. GNU General Public Licence. Version 1. February 1989. Free Software Foundation. 1989. Available from: https://www.gnu.org/licenses/licenses.en.html#GPL. Cited 29 August 2018.

30. Stallman R. The GNU-Manifesto. GNU Project–Free Software Foundation (FSF). 1985. Available from: https://www.gnu.org/gnu/manifesto.en.html. Cited 29 August 2018.

31. Michaels T. General Public License for Plant Germplasm: A Proposal by Tom Michaels. Bean Improvement Cooperative Conference. 1999.

32. Kipp M. Software and Seeds: Open Source Methods. First Monday. 2005. 10 (9).

33. Hope J. Biobazaar:The Open Source Revolution and Biotechnology. Harvard University Press. 2008.

34. Luby CH, Kloppenburg J, Michaels TE, Goldman IL. Enhancing Freedom to Operate for Plant Breeders and Farmers through Open Source Plant Breeding. Crop Science. 2015. Vol 55. 2481–2488.

35. Kloppenburg J. Impeding Dispossession, Enabling Repossession: Biological Open Source and the Recovery of Seed Sovereignty. Journal of Agrarian Change. 2010. 10, 367–388.

36. Luby CH, Goldman IL. Freeing Crop Genetics through the Open Source Seed Initiative. PLoS Biol. 2016. https://doi.org/10.1371/journal.pbio.1002441

37. Kotschi J, Kaiser G. Open-Source für Saatgut. Diskussionspapier. Agrecol—Association for AgriCulture and Ecology. 2012. Available from: http://www.agrecol.de/files2/open_source_studie_kotschi_kaiser.pdf Cited 29 August 2018.

38. CSA, Center for Sustainable Agriculture. Open Source Seed Systems. 2014. Available from http://csa-india.org/wp-content/uploads/2014/11/Open_Source_Seed_Systems_1.0.pdf Cited 29 August 2018.

39. HIVOS. Open Source Seeds. No year. Fact Sheet. Available from https://hivos.org/sites/default/files/open_source_seed_systems_factsheet.pdf. Cited 29 August 2018.

40. Kotschi J, Rapf K. Liberating seeds with an Open Source Seed Licence. Working paper. AGRECOL. 2016. Available from: www.agrecol.de/files/OSS_Licence_AGRECOL_eng.pdf. Cited 29 August 2018.

41. Apimondia. Press Release: Apimondia announces Open Source License to enable the Conservation and Sustainable Use of Bee Genetic Resources. World Beekeepers Association. 2016. Available from: http://www.apimondia.com/documents/gmo/apimondia_press_release_license.pdf. Cited 29 August 2018.

42. Bürgerliches Gesetzbuch (BGB), § 305. Bundesministerium für Justiz und Verbraucherschutz. (2017). Available from: http://www.gesetze-im-internet.de/bgb/BJNR001950896.html#BJNR001950896BJNG023401377. Cited 29 August 2018.

43. CBD, Convention on Biological Diversity. The Nagoya Protocol on Access and Benefit- Sharing. 2014. Available from: https://www.cbd.int/abs/text/articles. Cited 29 August 2018.

44. European Union. Regulation (EU) No 511/2014 of the European Parliament and of the Council of 16 April 2014 on compliance measures for users from the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization in the Union Text with EEA relevance. Official Journal of the European Union. 2014. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32014R0511. Cited 29 August 2018.

45. Agrecol. OpenSourceSeeds. 2017. Available from: https://www.opensourceseeds.org/en. Cited 29 August 2018.

46. Horneburg B. Participation, utilization and development of genetic resources in the Organic Outdoor Tomato Project. In: Goldringer I, Dawson J, Rey F, Vettoretti A, editors. Breeding for resilience: a strategy for organic and low-input farming systems? EUCARPIA, 2nd Conference of the "Organic and Low-Input Agriculture" Section. 2010. pp139-142.

47. Horneburg B, Becker HC. Selection for *Phytophthora* field resistance in the F_2 generation of organic outdoor tomatoes. Euphytica. 2011. 180:357–367.

48. Kotschi J and Wirz J. Who pays for seeds? Working paper. AGRECOL and Section for Agriculture Goetheanum. 2015.

6.10 Case Study 3

RESEARCH ARTICLE

Patent pledges, open IP, or patent pools? Developing taxonomies in the thicket of terminologies

Jonas Fabian Ehrnsperger*, Frank Tietze

Centre for Technology Management, Innovation and Intellectual Property Management (IIPM), Institute for Manufacturing, University of Cambridge, Cambridge, United Kingdom

* *jfe27@cam.ac.uk*

Abstract

Recently, a range of organisations, including car and consumer electronics manufacturers, have applied so-called patent pledges. A patent pledge is a publicly announced intervention by patent-owning entities ('pledgers') to out-license active patents to the restricted or unrestricted public free from or bound to certain conditions for a reasonable or no monetary compensation. Despite growing research to better understand this phenomenon, the underlying terminology remains contradictory. We apply an inductive research approach using qualitative coding to analyse 60 patent pledges made by 80 organisations. Based on this analysis, we propose a three-dimensional taxonomy that distinguishes eight types of patent pledges. Extending this taxonomy using case examples, we then propose a generalised patent licensing taxonomy. This second taxonomy can be used to distinguish patent licensing strategies, including other frequently used approaches, such as patent pools and cross-licenses. Finally, we use the patent pledge taxonomy to illustrate how patent owners change their licensing strategies over time and how it can support strategic decision processes within an organisation. We contribute to the field of patent management by building an ontology of patent pledges through proposing a definition and eight types. The patent licensing taxonomy enables organisations to devise and choose licensing strategies, and to illustrate licensing approaches of competitors, for instance.

Introduction

Existing theories explaining why and how organisations use Intellectual Property Rights (IPRs), such as those considering *revenue* and *profit dissipation effects* hardly explain why innovative firms such as Tesla Motors give away patents for free (i.e. without demanding royalty payments) [1]. While these effects and further theories play an important role in the licensing literature, the increasingly complex business environment and certain novel licensing approaches require the existing research to be adapted and extended.

The rise of open innovation and open source software caused some organisations to rethink long-established practices, including their attitude towards IP strategies [2– 4]. Subsequently, the announcement made by Tesla in 2014 triggered an ongoing debate inside and outside the academic world about what has been often labelled an *open IP strategy* more broadly or *patent pledge* specifically. While Tesla's patent pledge seems the mostly widely discussed pledge to date, it is by far not the only one. Other lesser-known examples for freely made available IPRs include the QR code technology. Even though such pledges have grown in frequency,

https://doi.org/10.1371/journal.pbio.3000023 Used with CC 4.0 license

they are far from new [5, 6]. Historic examples from the early 2000s include IBM's pledge not to assert 500 of its patents against the development, distribution and use of open-source software. Hence, the rise of patent pledges suggests a demand for more research.

This demand has been recognised by researchers, which resulted in some seminal works in the area of patent pledges [7]. The increasing research, however, has created ambiguity: academics and industry professionals are inconsistent in their use of terms such as patent pledges and open IP. For many, only IPRs that are free of charge and available to the unrestricted public constitute a patent pledge. Others adopt a broader view by considering those to be pledges where access to IPRs are available not for free, but for reduced licensing-fees. This inconsistency hinders the comparison of academic work and impedes communication between IP professionals. Furthermore, whilst many scholars use the notion of *IP pledges*, they generalise across all IPRs. However, since distinct IPRs protect different kinds of intellectual assets by law, it is too great a leap to assume that they all entail the same licensing strategies. For instance, it appears reasonable to assume that organisations apply different licensing strategies for patents (that protect technical inventions for a maximum of 20 years) and trademarks (that protect recognisable expressions that can theoretically be protected forever).

This paper addresses these problems by first proposing an inductively derived definition for patent pledges; second, by proposing a taxonomy of patent pledges consisting of eight mutually exclusive types; third, by proposing a patent licensing taxonomy that puts patent pledges into perspective to other common licensing types, such as cross-licenses and patent pools. The taxonomy of patent pledges emerges directly from patent licensing statements, whereas the patent licensing taxonomy constitutes an extension of the patent pledge taxonomy, and is validated using case studies. With these taxonomies, we aim to facilitate the distinction between different types of patent pledges and other patent licensing strategies by positioning them in relation to each other, making it possible to recognise differences as well as similarities. Hence, this paper contributes to the ontology for the patent licensing theory. As we focus only on patents, we provide a detailed overview of patent licensing that does not generalise across all types of IPRs.

This paper is structured as follows: section 2 summarises existing research on patent pledges contrasting existing definitions. Section 3 describes the data collection and qualitative coding process, which lay the groundwork for the taxonomy development. In section 4 we elaborate a definition of patent pledges and describe the patent pledge taxonomy and the patent licensing taxonomy. Section 5 discusses the findings, specifically focusing on changing licensing approaches of case examples within the context of the patent licensing taxonomy, and also emphasises limitations and suggestions for future research. Lastly, section 6 concludes by summarising the contributions of this paper.

Theoretical background

Patent pledges can be interpreted as a phenomenon that has only gained interest during the past decade. Despite its importance in both the legal and economic literature, it has not been formally studied until 2012 [6]. However, patent pledges do not appear to be a new phenomenon. Burnett provides examples of organisations that applied what he calls *'forfeiture actions'* dating back to the 1940s [5]. These actions exhibit similarities to what is currently labelled a *patent pledge*, including the reasonable price for the respective IPRs and its broad availability. Despite this early occurrence of patent pledges in the business environment, its focus in the literature remained scarce. In 1983, Allen became one of the first scholars to specifically investigate the free exchange of knowledge between firms [8]. Whilst not explicitly referring to the notion of patent pledges, he describes the *'free exchange of information about new techniques and plant designs among firms'* as a requirement for collective invention ([8], pg. 2).

https://doi.org/10.1371/journal.pbio.3000023

20 years later, Harhoff, Henkel, and Hippel followed up on this idea by defining '*the free revealing of information by a possessor as the granting of access to all interested agents without imposition of any direct payment*' ([9], pg. 1753).

After these early works, an increasing but small number of scholars turned to patent pledges. According to the definition respective scholars use, the modern research can be broadly divided into two 'camps'. On the one hand, some define patent pledges strictly as the access to patents without any monetary compensation. For example, Ziegler, Gassmann, and Friesike state that '*Patent release or give away for free means that in contrast to classic licensing and cross-licensing agreements, there is no contractual definition of compensation from the receiving end to the original patent holder*' ([10], pg. 19). Similar definitions are provided by Raasch, Herstatt, and Balka, Schultz and Urban, Alexy, George, and Salter, Asay, Sundaresan, Jena, and Nerkar and Contreras, Hall, and Helmers [11– 16]. Meanwhile, a second group widen the definition of patent pledges, including the access to patents on reasonable royalty-rates. For instance, Chander and Sunder broaden their definition of a *public domain* by adding the option to demand a nominal fee rather than no monetary compensation at all [17]. This idea is supported by Contreras and Contreras and Jacob [7, 18]. Chesbrough promotes this definition specifically in the context of Open Innovation [19].

Table 6.5 contrasts previous patent pledge definitions. It is important to note, however, that not all scholars specifically use the term patent pledges. In order to retain validity, we evaluated the content of definitions rather than the terminology itself.

Table 6.5 Existing patent pledge definitions.

Definitions including monetary compensation	Definitions excluding monetary compensation
'Resources for which legal rights to access and use for free (or for nominal sums) are held broadly.' [17], pg. 1338	*'Patent release or give away for free means that in contrast to classic licensing and cross-licensing agreements, there is no contractual definition of compensation from the receiving end to the original patent holder.'* [10], pg. 19
'These pledges encompass a wide range of technologies and firms : from promises by multinational corporations like IBM and Google not to assert patents against open-source software users; to commitments by developers of industry standards to grant licenses on terms that are fair, reasonable, and non-discriminatory (FRAND) …' [18], pg. 787	*'OSI [Open Source Innovation] is characterised by free revealing of information on a new design with the intention of collaborative development of a single design or a limited number of related designs for market or non-market exploitation.'* [11], pg. 2
'Pledge commitments fall into three general categories : (1) the primary commitment to license patents, either on royalty-free or FRAND terms, or not to assert patents at all …' [7], pg. 13	*'Patent pledges are promises by patent holders not to enforce their patents under certain conditions.'* [12], pg. 30
	'… selective revealing as the voluntary, purposeful, and irrevocable disclosure of specifically selected resources, usually knowledge based, which the firm could have otherwise kept proprietary, so that they become available to a large share or even all of the general public, including competitors.' [13], pg. 272

(*Continued*)

Table 6.5 Existing patent pledge definitions. (*Continued*)

Definitions including monetary compensation	Definitions excluding monetary compensation
	'Parties are increasingly engaging in "patent pledging," a phenomenon where parties voluntarily commit to limit enforcement of their patent rights.' [14], pg. 261
	'An open IP strategy is any strategy that allows external inventors and firms the use of technology developed by the focal firm without any financial or cross-licensing obligation.' [15], pg. n.a.
	'…under a pledge model, patent assets are retained by their owners, who continue to incur maintenance and other fees, but the offensive use of such patents is significantly curtailed.' [16], pg. 1

https://doi.org/10.1371/journal.pone.0221411.t001

The contrasting definitions indicate that different types of patent pledges may exist. Contreras and Jacob distinguish between the price and its underlying calculation. They describe three broad categories [7]: *Primary Access Commitments, Secondary Royalty Commitments* and *Non-royalty Commitments* of patent owners. While this empirically derived categorisation is, to our knowledge, the first attempt towards a taxonomy of patent pledges, it also includes general promises that do not aim to facilitate access to IPRs, such as the pledge not to transfer patents to non-practicing entities (NPEs) [20]. Also, the authors' categorisation involves the commitment to facilitate prior-art searches, and this is not directly related to patent licensing.

It is reasonable to conclude that academics previously supported two patent pledge definitions: (i) pledges that demand a reasonable fee for the usage, and (ii) pledges that do not demand any monetary compensation at all. This inconsistency impedes the comparison of different results, leading to a contorted understanding of patent pledges.

Research approach

Data collection

Similar to the analysis of Lerner, Strojwas and Tirole, we collected and coded official licensing agreements [21]. First, we had to decide on which patent initiatives to include in this study, i.e. which could be classified as patent pledges. As a starting point, we used the patent pledge dataset published by Contreras [22]. This dataset is, to our knowledge, the only publicly available dataset that includes patent pledges. The dataset lists 178 pledges and includes accompanying documents in which organisations pledge to apply a range of specific IP practices. However, the entries on this list are organisation-specific: every organisation that pledges a specific IP practice is counted as one entry (apart from the *Open Invention Network*). In order to avoid the inclusion of identical pledges, we consolidated the organisations that took part in the same initiative—counting pledges rather than individual organisations. To further filter the data, we excluded 13 entries: (1) the facilitation for prior-art searches ('The Clearing House', Microsoft, Yahoo, SAS), (2) the promise to

enable/improve the community review of IPRs (IBM), (3) entries with vague formulations (Novell's patent policy from 2014, Allergan's social contract with patients), (4) the promise not to sell IPRs to non-practicing entities (Verizon, Cisco Systems), (5) entries with a lack of information (John Gilmore, '*Patent Licensing Principles*' by Conversant), (6) an entry with the strict restriction to qualified customers (Microsoft's Azure IP Advantage programme), (7) the mere statement of availability of licenses without the specification of standardized or reasonable terms (Microsoft). Furthermore, we added *The GreenXchange initiative, the OpenPOWER foundation* and patents relating to the QR-Code technology, because they specifically address the reasonable or free availability of patents based on standardised terms.

In total, our dataset comprised 60 patent pledges across 80 organisations (the *Open Invention Network*, the *OpenPOWER foundation*, the *Eco-Patent Commons, and* the *GreenXchange* are counted as one organisation; Ericsson, Sony, Sony Ericsson, Nokia, Siemens, Nokia Siemens Networks are counted separately.) The full dataset includes about 260 pages of DIN-A4 PDF-files. Only in the cases of GreenXchange and the Eco-Patent Commons (both no longer provide an active website) do we rely on empirical data and descriptions provided elsewhere [16, 23]. The complete list with information about number, type and area of technology is provided in S1 Table. It is important to note, however, that in some cases several patent pledges are announced in one initiative, which results in fewer dataset files than counted patent pledges. For instance, the patent pledge 'Microsoft 03.12.2003' is one PDF-file in which the firm announces three different patent pledges.

Data analysis

We coded the data in two cycles using NVivo software. While qualitative coding can be conducted manually, we used the software to facilitate the process of deleting, subsuming and recoding initial codes, and to facilitate the analysis. We did, however, find the raw data straightforward and consistent—we were only required to occasionally change codes. In a first coding cycle, we followed the qualitative analysis approach of conventional content analysis. We therefore avoided preconceived categories by allowing common themes to emerge from the raw data [24, 25]. Through initial coding we remained open to new interpretations by limiting the influence of our own subjective judgement [26, 27]. However, as suggested in the literature, we noted tentative ideas for themes that naturally arise through the human inclination of pattern-seeking [27]. In a second coding cycle, specifically through pattern coding, we revisited the data to reorganise and adapt initial concepts [27]. We structured the emerging concepts by summarising them into 1st order codes, 2nd order codes and aggregate dimensions, as suggested by [28, 29]. Three dimensions emerged from the coding: *Accessibility, Compensation* and *Conditions* (see Fig 6.1).

Taxonomy development

The terms *taxonomy* and *typology* are often used interchangeably in the literature [30]. A difference between these terms is that the former consists of mutually exclusive sets while the latter relates to a set of ideal types. Crucially, only taxonomies provide key decisions that allow the classification of different organisations [31]. We avoid the common problem of summarising existing research in typology-based theory-building by using data that has not been analysed [32]. Since the proposed dimensions emerged directly from the dataset, we are confident that they are representative. Furthermore, they fulfil the criteria of *parsimony*, which Nickerson, Varshney and Muntermann describe as the manageable number of dimensions to prevent overextension [30]. At the same time, however, the dimensions allow for a clear distinction from each other. All elements of our dataset express forms of *Accessibility, Compensation* and *Conditions*, and this makes the taxonomy comprehensive [30].

Results

Definition of patent pledges

We propose the following definition of patent pledges that consists of the three elements that we derived from the content analysis of our dataset: *Accessibility, Compensation* and *Conditions*. It should be noted that we only consider active patents and pledges that are publicly announced (i.e. the unrestricted public can access the information).

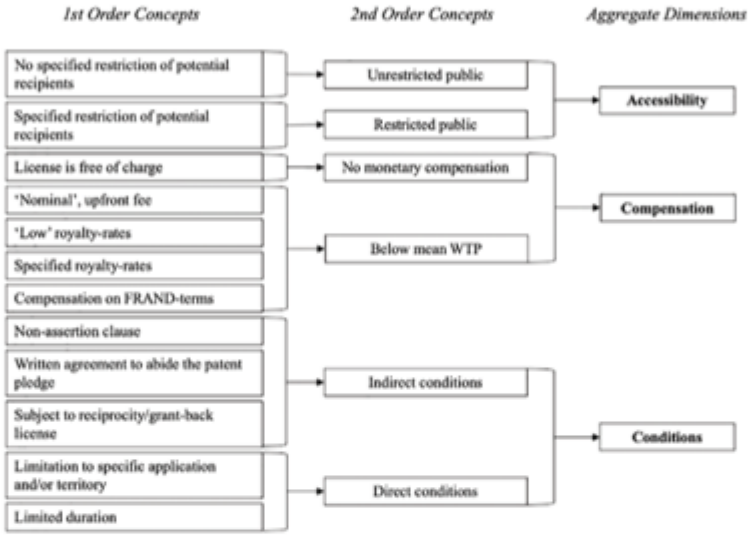


Fig 6.1 Coding process and emerging dimensions.

https://doi.org/10.1371/journal.pone.0221411.g001

> *'A patent pledge is a publicly announced intervention by patent owning entities ('pledgers') to out-license active patents to the restricted or unrestricted public free from or bound to certain conditions for a reasonable or no monetary compensation using standardized written or social contracts.'*

The restriction to active patents contrasts with patent donations, which generally refer to abandoned rights. The term *reasonable* in this context refers to fair, reasonable and non-discriminatory terms (FRAND-terms). We follow the definition of FRAND-terms described by [33]. There exist different approaches to calculate FRAND-terms. For instance, they can be economically determined by describing a hypothetical scenario of negotiations between the licensor and the licensee. A reasonable royalty-rate falls somewhere between the patent holder's minimum fee they are willing to accept and the would-be infringer's maximum willingness to pay (WTP) [34]. We follow the definition of WTP, or reservation price, described in the literature [35– 37]. A reasonable royalty-rate is considered lower than the maximum WTP by the potential licensee, which is an important characteristic of our second taxonomy: the patent licensing taxonomy. The maximum WTP can be averaged among the elements of interest, which results in the *mean WTP*. Many studies apply mathematical and statistical methods to calculate the mean WTP for a specific business area [38, 39]. This value serves as a directive for a rough estimate of common royalty-rates in a specific industry. For example, the average royalty-rate in the automotive industry is given as 4.7% of the sales price, whereas in pharmaceuticals it is estimated to be 7.0% [40]. While the proposed taxonomies attempt to be general rather than industry-specific, we leave aside the calculation of mean WTP values for distinct business areas.

Patent pledge taxonomy

For the taxonomy development, we applied a dichotomous scale to each of the three aggregate dimensions. This results in a taxonomy with eight types. Subsequently, we describe each dimension and the values it can take.

Accessibility is a measure for the potential recipients of the respective patents. Our analysis shows that patent pledges either address a large number of third parties or the public in general. We therefore suggest dividing this dimension into the two categories of *Restricted Public* and *Unrestricted Public*. Significantly, the *Restricted Public* category is only concerned with the a priori restriction to specified licensees and must not be confused with the dimension *Conditions* described below. This can be illustrated by the following example. Toyota pledges the availability of specified patents only '*to automakers who will produce and sell fuel cell vehicles, as well as to fuel cell parts suppliers and energy companies*'. This patent pledge restricts the number of potential licensees from the outset. On the other hand, IBM's patent pledge from January 2005 does not restrict licensees a priori, but the use of their IPRs is subject to a specified condition: '*IBM hereby commits not to assert any of the 500 U.S. patents listed below, as well as all counterparts of these patents issued in other countries, against the development, use or distribution of Open Source Software*.' This distinction is essential, since in the former example only specified third parties can make use of the patent pledge (i.e. restricted public). Whilst in the latter, anyone—subject to specific conditions—can access the patents (i.e. unrestricted public).

Compensation refers to the direct monetary compensation the pledger demands in exchange for the license. Our analysis reveals that patent pledges are either made in return for a *reasonable compensation* or *free of charge*. As outlined above, a reasonable compensation in the literature is considered to be below the mean WTP. Therefore, we suggest distinguishing this dimension in the two categories *Below mean WTP* and *None*.

Conditions is a measure for any usage-related condition for the respective patents. Our analysis shows that 42 out of 60 patent pledges are subject to explicitly mentioned conditions. The coding process revealed five non-exclusive categories of conditions listed in Table 6.6 below. We further distinguish the conditions according to the influence they have on the respective patent-usage. We suggest to summarise conditions that restrict the usage in either space (physical or technological) or time as **direct conditions**. In our sample, direct conditions are the restriction to a specific technology field/territory or a time limitation for the license. In contrast, **indirect conditions** do not affect the patent-usage in either space or time, such as the promise not to assert patents against the licensor (non-assertion clause).

Table 6.6 Patent pledge conditions.

	Condition	Example
Direct conditions	**Limitation to specific applications and/or territory** (in 25 out of 60 pledges)	*'The policy also broadens Microsoft's commitment to provide the academic community with IP under royalty-free terms for noncommercial use.'* Source: Microsoft pledge, March 2003
	Limited duration (in 3 out of 60 pledges)	*'Patents related to fuel cell vehicles will be available for royalty-free licenses until the end of 2020.'* Source: Toyota pledge, 2015

(*Continued*)

https://doi.org/10.1371/journal.pbio.3000023

Table 6.6 Patent pledge conditions. (*Continued*)

	Condition	Example
Indirect conditions	**Non-assertion clause** (in 21 out of 60 pledges)	*'A party is "acting in good faith" for so long as such party and its related or affiliated companies have not : asserted, helped others assert or had a financial stake in any assertion of (i) any patent or other intellectual property right against Tesla or (ii) any patent right against a third party for its use of technologies relating to electric vehicles or related equipment …'* Source: Tesla Motors pledge from 2014, Status: February 2019
	Subject to reciprocity/ grant-back license (in 10 out of 60 pledges)	*' Qualcomm has had a long standing policy of broadly offering to license its standards essential patents for CDMA-based telecommunications standards on terms and conditions that are fair, reasonable, and free from unfair discrimination (FRAND), subject to reciprocity.'* Source: Qualcomm pledge, 2008
	Written agreement to abide the patent pledge (in 1 out of 60 pledges)	*'Thus, Google will require any person or entity to whom it sells or transfers any of the Pledged Patents to agree, in writing, to abide by the Pledge and to place a similar requirement on any subsequent transferees to do the same.'* Source: Google open patent non-assertion pledge. Status: February 2019

https://doi.org/10.1371/journal.pone.0221411.t002

Fig 6.2 shows the proposed patent pledge taxonomy. Subsequently, we describe each of the eight pledge types by providing an example. In this description, we do not add the *Conditions* dimension, since the conditions described in Table 6.6 can be equally applied to all types of patent pledges.

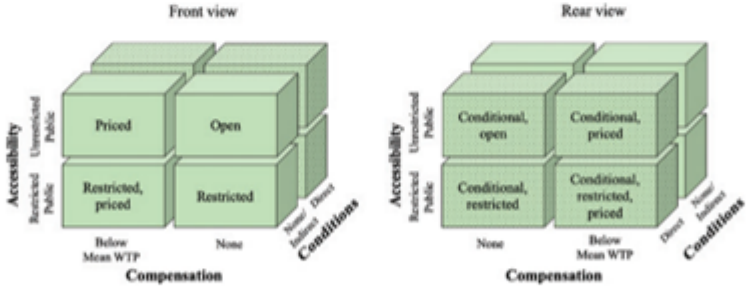


Fig 6.2 Patent pledge taxonomy.

A **restricted, priced patent pledge** is a pledge that makes patents available to the restricted public in exchange for a reasonable price. An example for this type is iBiquity's pledge from 2005, in which the organisation commits to license patents on FRAND-terms for *'someone who is skilled in the art to manufacture NRSC-5 compliant transmission devices'*. The firm does not provide a description of these required skills, which leaves it open for interpretation.

https://doi.org/10.1371/journal.pbio.3000023

A **priced patent pledge** is a pledge made from patent owning entities by announcing the general willingness to out-license on FRAND-terms. Significantly, the compensation *Below Mean WTP* must be equally available to all interested parties. An example is the patent pledge from NTT DoCoMo et al. from 2002, in which the companies pledge the availability of licenses relating to the W-CDMA technology to the unrestricted public for '*a cumulative royalty-rate below 5%*'.

A **restricted patent pledge** describes the free availability of licenses for the restricted public. The merger thepatentpledge.org is an example of this type in which 35 organisations pledge not to assert their software patents against firms that employ less than 25 people [41]. Another example is Toyota's patent pledge from 2015, in which the firm announced that patents relating to fuel-cell technology would be freely available to '*automakers who will produce and sell fuel cell vehicles, as well as to fuel cell parts suppliers and energy companies who establish and operate fuelling stations*'.

An **open patent pledge** constitutes patents that are freely available to the unrestricted public. Tesla Motors' pledge from 2014, for instance, is of this type. Even though there has been much criticism of its initial inaccuracy, Tesla has since provided an explanation of terms and conditions on its website: '*Tesla irrevocably pledges that it will not initiate a lawsuit against any party for infringing a Tesla Patent through activity relating to electric vehicles or related equipment for so long as such party is acting in good faith*' [42]. In this case, the pledge is restricted to technologies relating to electric vehicles and is also condition to a non-assertion clause.

Following this, we provide data regarding the distribution of our sample across the patent pledge types and technology areas (see S1 Table for an individual breakdown). 19 pledges (32%) are classified as *conditional, open*, 12 as *conditional, priced (20%)*, 11 as *conditional, restricted (18%)*, 9 as *priced (15%)*, 5 as *open* (8%), 3 as *restricted* (5%), and 1 as *restricted, priced* (2%). Our sample did not show patent pledges that can be classified as *conditional, restricted, priced*. 47 out of 60 patent pledges (78%) relate to Information and Communication Technology (ICT). 6 pledges (10%) can be allocated to technologies that aim at benefiting the environment; 4 (7%) can be assigned to biotechnology and genetics. 3 patent pledges are made by automobile manufacturers, 2 (3%) relate to electric vehicles (Tesla Motors and Ford) and 1 (2%) to fuel-cell vehicles (Toyota).

The patent licensing taxonomy

Academics emphasise that dimensions of reliable taxonomies must be extendable [30]. By examining the patent pledge taxonomy, we find that especially the dimensions *Accessibility* and *Compensation* allow for an extension on one side of the scale. Specifically, there is the possibility that organisations allow very few or no third parties to access their patents. Similarly, they can demand the mean WTP or a higher price as compensation. In contrast, we find no use in extending the dimension *Conditions*, because we have already included the *None* category in the patent pledge taxonomy, and we cannot find more restrictive conditions than the direct restrictions. Therefore, we propose to extend the two dimensions, *Accessibility* and *Compensation*, to categorise patent licensing approaches that are beyond the definition of patent pledges, which then leads to a general patent licensing taxonomy (see Fig 6.3). In the following section, we describe purposefully, non-randomly selected examples of organisations to substantiate the resulting types [43].

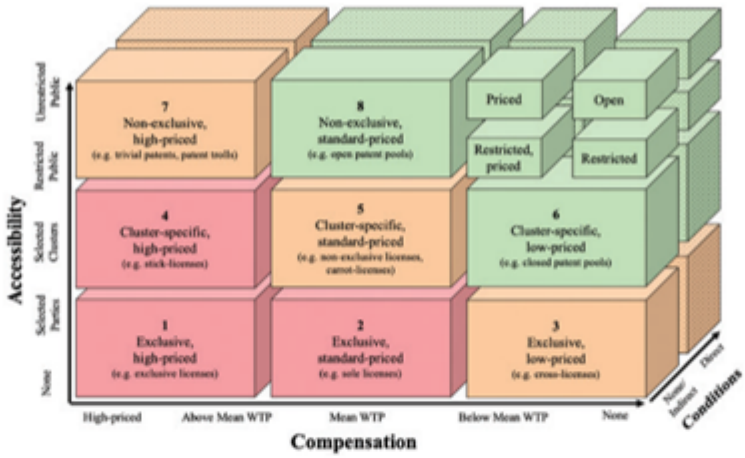


Fig 6.3 The patent licensing taxonomy.

https://doi.org/10.1371/journal.pone.0221411.g003

As Fig 6.3 shows, the dimension *Compensation* has been extended to cover the categories *High-Priced, Above Mean WTP*, and *Mean WTP*. It now ranges from high-priced to free of charge. The relationship between FRAND-terms and the WTP, which is the maximal price at which a customer would buy a product or service, has been summarised above [35]. For specific technology areas, certain values for the mean WTP exist. Therefore, the patent licensing taxonomy should be used with regards to specific technologies.

The dimension *Accessibility* has now been extended to cover three additional categories, *Selected Clusters, Selected Parties*, and *None*. While the term *Cluster* often refers to the geographical location of firms, we use the term to address a conglomerate of entities defined by the licensor [44]. Therefore, a cluster in the context of the patent licensing taxonomy is a number that falls somewhere between *Selected Parties* and *Restricted Public*. Generally speaking, when an organisation specifies a cluster, it sacrifices some control over the number of potential recipients by addressing a specific group of organisations that can change without the influence of the licensor. This stays in contrast to the category *Selected Parties*, in which the licensor addresses specifically named organisations while retaining full control over the number. However, the number of recipients in a cluster is intended to be smaller than the number referring to the restricted public.

The patent pledge taxonomy with its eight distinct types proposed in the previous section constitutes the upper right cornerstone of the patent licensing taxonomy. The distinction between licensing types regarding their *Accessibility, Compensation* and *Conditions* leads to a distinction regarding their *openness* and *exclusiveness*. For instance, *cluster-specific, low-priced licensing strategies* in the patent licensing taxonomy are more open with regards to their *Accessibility* and *Compensation* than *exclusive, standard-priced licensing strategies*. In contrast, they are more closed regarding their *Accessibility* than patent pledges. Generally, closed approaches in the patent licensing taxonomy are types 1, 2, and 4. Intermediate approaches are types 3, 5, and 7, whilst open approaches are types 6, 8, and all patent pledges. Henceforth, we provide examples for each of the eight remaining types, numbered 1–8, drawing on prior literature.

1 Exclusive, high-priced licensing strategies. *Exclusive, high-priced licensing strategies* represent a proprietary approach to an organisation's patents. Patent owners that employ this approach only share their patents for a high price to carefully selected parties, if at all. Some organisations believe that the exclusion of other parties from their patented invention is the most valuable strategy. We argue that these patent owners would out-license their patents if the price was high enough. However, in many cases they would charge

such a high price that interested parties look for other ways to enter the market. This can result in what the academic literature often calls a *monopoly* [45, 46].

Examples for *exclusive, high-priced licensing strategies* frequently occur in the pharmaceutical industry. While some patent owners of pharmaceutical drugs license their patents on a high price, most of them refuse to grant a license when there exist no compulsory licensing laws in the respective country [47, 48]. Examples in other industries include Polaroid and Philips. Polaroid excluded its competitor Kodak from the instant camera industry in 1981, and Philips used patents to maintain a monopoly over a specific shaving technology [40, 49]. Generally, exclusive licenses fall under this category.

2 Exclusive, standard-priced licensing strategies. *Exclusive, standard-priced licensing strategies* are characterised by a monetary compensation that ranges closely around the mean WTP. Analogous to type 1, the patent owner only allows a few parties to access the respective patents.

Exclusive licenses for different territories, which only allow one party the usage of patents in a specific area, are an example for this type. For instance, the initial polyester patent was exclusively licensed to Du Pont for exploitation in the U.S., while Imperial Chemical Industries (ICI) held the exclusive license for the rest of the world [50]. The territorial restrictions in this example caused the price to go down from high-priced to standard-priced. Licenses in which only one licensee and the licensor are allowed to use the patents are commonly known as *sole licenses*, which are also a form of this type [40].

3 Exclusive, low-priced licensing strategies. *Exclusive, low-priced licensing strategies* are accessible to selected parties at a price that is below the mean WTP. Organisations that operate this licensing type carefully select who they allow to use their patents, but the respective patents are also reasonably priced or free of charge.

Hewlett-Packard (HP) is an example of an organisation that uses this licensing strategy. With its wide range of products and a large number of competitors and suppliers, many of HP's licensing activities in the past did not aim to generate revenue. Rather, its goal was to establish long-term business partnerships with selected parties that built upon low royalty-rates [51]. Generally, open innovation collaborations with common IP-transfers are a further example for *exclusive, low-priced licensing strategies* [2]. Lastly, cross-licensing agreements with or without additional royalty-payments also fall into this category. For instance, at one point both Intel and IBM entered a cross-licensing agreement, which automatically licensed a major part of the respective firm's patents to the other party [52].

4 Cluster-specific, high-priced licensing strategies. *Cluster-specific, high-priced licensing strategies* are characterised by licensing-arrangements that offer patents to a cluster of organisations at a price that is above the mean WTP.

In this context, HP provides another example. HP owned patents that were not necessarily of strategic importance to them but that were essential to other organisations of a specific cluster [51]. Here, the firm used out-licensing to generate additional revenue. Yet to keep the price high, the license had not been offered to the restricted or unrestricted public to allow potential licensees to keep a competitive advantage. Hence, HP focused on the monetisation of patents rather than the development of long-term business partnerships which have been described in type number 3. This strategy thus seems to be particularly interesting for patents that are of strategic importance to other organisations. When such organisations involuntarily need to license an invention, the resulting licenses are often referred to as *stick-licenses* [40]. Organisations active in the technology market serve as another example. Qualcomm faced major competition for producing handsets embodying its Code Division Multiple Accessibility (CDMA) technology. To increase revenue, Qualcomm focused on out-licensing to specified clusters rather than producing the technology itself [53].

5 Cluster-specific, standard-priced licensing strategies. *Cluster-specific, standard-priced licensing strategies* refer to licensing approaches with patents that are made accessible to a cluster of third parties at a price that is close to the mean WTP.

This licensing type describes an approach in which organisations license their patents to third parties on a common price for this specific cluster [53]. For instance, the Cambridge University spin-off Cambridge Display Technologies (CDT) developed light-emitting plastics and initially tried to manufacture its products on its own [54]. Close to insolvency, CDT changed its business model and out-licensed the technology to entrenched manufacturers on royalty-rates that were common for this industry. CDT needed fast revenues and had no core technology that was essential for other firms. To survive, the firm offered so-called *carrot-licenses* in an attempt to incentivise others to buy a license [40].

6 Cluster-specific, low-priced licensing strategies. *Cluster-specific, low-priced licensing strategies* differ from patent pledges in the sense that licensors offer the respective patents to a specific cluster only.

Closed patent pools are an example for this licensing type. A closed patent pool consists of at least three parties and allows its members to license patents from other members either for free or at reasonable costs [21, 55, 56]. For instance, the License on Transfer Network (LOT-Network) offers its members both free and reasonable licenses depending on the size of the licensee. LOT was founded by Google, Canon and RedHat and grants its members a license to about 1.2 million IPRs in case those rights are transferred to an NPE (indirect condition). Interested parties must join the LOT-Network and pay an annual fee for the membership. This fee is based on the annual revenue of each firm and ranges from 0 US-Dollars (for firms that generate less than 25 million US-Dollars annually) to a maximum fee of 20.000 US-Dollars (for firms that generate more than 1 billion US-Dollars annually). Furthermore, a limited number of start-ups receive access to royalty-free licenses of three patents of their choice [57]. Importantly, only LOT-members can benefit from these patents.

7 Non-exclusive, high-priced licensing strategies. *Non-exclusive, high-priced licensing strategies* refer to the licensing of patents that are offered to the restricted or unrestricted public for a price that exceeds the mean WTP.

The literature suggests that trivial patents which cover multiple constructive technologies can force entire industries into paying high royalty-fees, as the example of Amazon's 1-click patent from 1999 shows [52, 58, 59]. Examples also include NPEs (often referred to as *Patent Trolls*). NPEs out-license patents to third parties under the threat of litigation if the infringer refuses to buy a license. The threat of an injunction also influences the compensation [58–60]. However, as Chesbrough concludes, NPEs can take many forms and seldom use a consistent business model [2]. In sum, *non-exclusive, high-priced licensing strategies* can occur as a consequence of trivial patents or due to patent owners with primarily monetary motives that seek to enforce royalty-payments.

8 Non-exclusive, standard-priced licensing strategies. *Non-exclusive, standard-priced licensing strategies* are characterised by patents that are available to the restricted or unrestricted public at a price that is close to the mean WTP.

Open patent pools, in contrast to closed patent pools described in type number 6, constitute an example for this licensing approach. Open patent pools provide every interested third party with patent-packages at standard royalty-rates [21, 54, 56]. This is different from closed patent pools in the sense that an open patent pool does not require third parties to become a member. For instance, the patented technologies relating to the Moving Pictures Expert Group 2 (MPEG 2), the 3G Mobile Communication Standards and the Digital Versatile Disk (DVD) were being made available to the unrestricted public at standard royalty-rates through open patent pools [54, 56, 61]. This type includes IPRs that constitute a standard (standard-essential patents or SEPs) and must be made available to the unrestricted public on FRAND-terms. An important characteristic is that they must be offered to the restricted or unrestricted public. Compared to type number 6, this openness often comes at a higher price.

Discussion

This section will discuss how changes in licensing strategies can be displayed in the patent licensing taxonomy and how the taxonomy can help organisations during strategic decision processes, such as deciding between competing technologies. Furthermore, we will address the limitations of this study and give implications for future research.

Transitions through the patent licensing taxonomy over time

The examples provided in the previous section may suggest that organisations adopt only one licensing type. However, examples show that organisations adapt their licensing strategies over time, and thus transition from one type to another. Furthermore, it is obvious that organisations diversify their licensing strategies for different technologies. In one example, BP Chemicals licensed patents relating to acetic acid only very selectively, whereas it tried to license its patents in the area of polyethylene aggressively [53]. Subsequently, we will describe some case examples for specific changes in openness of licensing and illustrate them in Fig 6.4.

Fig 6.4 Exemplary paths through the patent licensing taxonomy.

https://doi.org/10.1371/journal.pone.0221411.g004

Transitions towards openness. Transitions from exclusiveness towards openness in patent licensing constitute one possibility to adapt licensing approaches over time. For instance, Du Pont held a monopoly for moisture-proof cellophane in the 1930s and allowed only one other company to use the technology under a high-priced license. During that time, Du Pont's licensing approach was an *exclusive, high-priced licensing strategy*. After its competitor Dow Chemicals introduced a new packaging material, Du Pont offered licenses to parties that had previously, without success, requested a license for cellophane. The new licenses had been offered on very reasonable terms in order to enable a quick start and deter Dow Chemical's entry into this specific market [50]. These new licensing-offers marked a transition of Du Pont's licensing strategy from *exclusive, high-priced* to *cluster-specific, low-priced licensing strategies*. Another example is Union Carbide. A major producer of ethylene glycol, the firm initially kept its own production process for glycol secret—these were *exclusive, high-priced licensing strategies*. Faced by increasing competition, Union Carbide started to out-license the process selectively and shifted their licensing strategy to *exclusive, standard-priced licensing strategies* [62].

Organisations do not necessarily adapt licensing strategies voluntarily. In the first 30 years after its foundation in 1885, AT&T used its patents in an exclusive manner to establish itself in the market. Forced by an antitrust consent decree in 1956, the firm was legally required to license its patents on reasonable royalty-rates below market value [51]. After the consent decree came to an end, AT&T returned to a more closed licensing approach, albeit not as closed as in its formative years.

Transitions towards exclusiveness. Organisations can also start with an open approach and adopt a more exclusive licensing approach over time. A reason for this might be the attempt to initially accelerate diffusion and adoption-rates of a patented technology. Once the technology is established, the licensor can return to a proprietary strategy and recover competitive advantages. For example, Toyota's patent pledge for the royalty-free use of thousands of patents relating to fuel cell vehicles is only valid until 2020. After that, it is possible that the car manufacturer could demand royalty-fees for the use of its patents, thus causing a shift to a more exclusive approach. Over 30 years ago, a similar approach has been described in the context of pricing strategies [63]. *Penetration pricing* is a pricing strategy used in the early stages of a new technology to incentivise as many adopters as possible. Once consumers rely on this adopted technology, the organisation can regain profits by gradually demanding higher prices at a later stage. This might also explain why firms that apply patent pledges do not simply abandon their patents. If they gave up the protection, they could neither demand licensing-fees nor seek injunction at a later stage. Another example of a firm that shifted from open to more exclusive licensing strategies is Nippon Shokubai with its process for producing acrylic acid in the 1970s. As the company changed its business model from licensing patents to the production of acrylic acid, it tightened its attitude by licensing only to markets that it could not supply itself [62].

Transitions along single dimensions. Organisations often adapt royalty-rates for their patents to changing competitive conditions and the altered value of its underlying technology [51]. They charge different rates for different licensees or, if permitted in the license agreement, they can change royalty-rates for the same licensee [51]. In one case, the US-firm Alliacense forces third parties into license agreements under the threat of litigations and thus qualifies as a NPE. As some scholars have pointed out, one particularity of this firm is that it offers early licensees' lower royalty-rates to incentivise rapid license agreements [58]. If an influential company first agrees to the reduced license, others are urged to follow at a higher price [58]. This licensing strategy constitutes the transition from *non-exclusive, standard-priced* to *non-exclusive, high-priced licensing strategies*. Similarly, patent owners can change the number of potential recipients and can introduce or eliminate conditions for the patent-usage.

Further discussions and future research

Our findings show that the majority of patent pledges relate to ICT, which is in line with earlier findings [14, 20]. This strengthens the connection between patent pledges, open innovation and open source software [64]. However, since the line between IP-protection and open source software is vague and also country-specific, future research should address this issue by specifically incorporating open source software into the patent licensing taxonomy.

Regarding the distribution of patent pledges, it is noteworthy that *conditional, restricted, priced patent pledges* did not occur in our sample. This could be a consequence of an organisation's intention that, when it decides to out-license patents to the restricted public in return for monetary compensation, it does not want to exclude potential licensees by imposing further conditions. While we are confident that both our taxonomies depict practical strategies that are actually used, our results are not definitive. Specifically, we aim to collect further patent pledge data to fill the gap of *conditional, restricted, priced patent pledges*. Future research should also look towards a more detailed investigation of individual patent pledge types, including the motives behind them and how these relate to specific business models.

Our second taxonomy, the patent licensing taxonomy, can be used as a strategic tool to support certain decisions within an organisation. For instance, when an organisation faces the decision between two competing technologies, it can compare respective licenses for each technology in the patent licensing taxonomy. This provides a profile of the technologies from the perspective of existing licenses, while the strategic importance of individual licenses can be visualised using different shape-sizes. As a result, the organisation gains an overview of the availability and distribution of patents across different owners, which ultimately influences the adoption decision.

Furthermore, the patent licensing taxonomy can uncover connections between firms that apply different business models. For example, whereas Alliacense and Nippon Shokubai are organisations with very different approaches, they both transitioned to a more exclusive licensing approach. The patent licensing taxonomy is a tool that allows for the illustration of multiple licensing strategies within the direct business environment of an organisation. We are aware that the categories of our taxonomies are debatable. To make this illustration more precise and transparent, prospective research should focus on ways to objectively quantify the categories. However, we emphasise that the taxonomies are especially useful when used to compare units of interest. It is easier to position organisations when looking at openness relative to each other, rather than in absolute terms.

Conclusion

Patent pledges are a phenomenon that has recently gained attention in both the day-to-day business of organisations and in the academic literature. A review of the existing research has shown that the term patent pledges still lacks a clear definition. We believe this is because the consideration of different types of patent pledges has been, to a large extent, overlooked. It is not just patent pledges but most patent licensing approaches that suffer from a lack of coherent definitions, which prohibits a clear distinction between licensing strategies and the comparison of academic research.

In this paper, we have proposed an inductively derived definition of patent pledges, and therefore contribute to the ontology of IP licensing. Furthermore, we have developed two taxonomies, one for patent pledges, and one for patent licensing in general. Both allow for a better distinction of existing licensing strategies in the current business environment and provide a coherent terminology. The patent pledge taxonomy classifies eight distinct types and enables a differentiation between licensing approaches that, until now, have been roughly termed *IP Pledges*. By extending two dimensions of this first taxonomy, we have created the patent licensing taxonomy. In addition to the facilitated distinction between common licensing approaches, the patent licensing taxonomy enables organisations to illustrate and compare the licensing strategies within their business environment, including those of their direct competitors. Both taxonomies can help managers to make informed decisions in the increasingly complex world of patent licensing.

Supporting information

S1 Table. Patent pledge dataset—Types and technology areas.
(DOCX)

Acknowledgments

We would like to thank Jorge Contreras for creating and sharing the list of patent pledges and for his invaluable feedback on an earlier draft of this paper. We thank Ove Granstrand for his support in the taxonomy-development and participants of the Open IP and Innovation and IP Management (IIPM) research seminars at the University of Cambridge. This research project is funded by the Engineering and Physical Sciences Research Council (EPSRC) and the Research and Development Management Association (RADMA). We

presented the patent pledge taxonomy at the 2019 International Conference of the IEEE Technology and Engineering Management Society in Atlanta, Georgia (USA).

Author Contributions

Conceptualization: Jonas Fabian Ehrnsperger.
Data curation: Jonas Fabian Ehrnsperger.
Formal analysis: Jonas Fabian Ehrnsperger.
Investigation: Jonas Fabian Ehrnsperger.
Methodology: Jonas Fabian Ehrnsperger, Frank Tietze.
Project administration: Jonas Fabian Ehrnsperger.
Supervision: Frank Tietze.
Validation: Jonas Fabian Ehrnsperger.
Visualization: Jonas Fabian Ehrnsperger.
Writing – original draft: Jonas Fabian Ehrnsperger.
Writing – review & editing: Jonas Fabian Ehrnsperger.

References

1. Arora A. and Fosfuri A., (2003), "Licensing the market for technology," *Journal of Economic Behavior & Organization*, vol. 52, no. 2, pp. 277–295.
2. Chesbrough H. W., *Open Innovation*: *The New Imperative for Creating and Profiting from Technology*. Harvard Business School Press, 2006.
3. Hippel E. and Krogh G., (2003), "Open Source Software and the 'Private-Collective' Innovation Model: Issues for Organization Science," *Organization Science*, vol. 14, no. 2, pp. 209–223.
4. Lerner J. and Tirole J., (2002), "Some Simple Economics of Open Source," *The Journal of Industrial Economics*, vol. 50, no. 2, pp. 197–234.
5. Burnett J. M., (2011), "The Host´s Dilemma: Strategic forfeiture in platform markets for informational goods," *Harvard Law Review*, vol. 124, no. 8, pp. 1861–1938.
6. Contreras, J. L., (2018), "The Evolving Patent Information Landscape," Centre for International Governance Innovation (CIGI), no. 166.
7. Contreras J. L. and Jacob, Meredith, *Patent Pledges*: *Global Perspectives on Patent Law's Private Ordering Frontier*. Edward Elgar Pub., 2017.
8. Allen R. C., (1983), "Collective Invention," *Journal of Economic Behavior and Organization*, vol. 4, pp. 1–24.
9. Harhoff D., Henkel J., and Hippel E., (2003), "Profiting from voluntary information spillovers: How users benefit by freely revealing their innovations," *Research Policy*, vol. 32, no. 10, pp. 1753–1769.
10. Ziegler N., Gassmann O., and Friesike S., (2014), "Why do firms give away their patents for free?," *World Patent Information*, vol. 37, pp. 19–25.
11. Raasch C., Herstatt C., and Balka K., (2009), "On the open design of tangible goods," R&D Management, vol. 39, no. 4, pp. 382–393.

12. Schultz J. and Urban J. M., (2012), "Protecting Open Innovation: The Defensive Patent License as a New Approach to Patent Threats, Transaction Costs, and Tactical Disarmament," *Harvard Journal of Law & Technology*, vol. 26, no. 1, pp. 1–61.

13. Alexy O., George G., and Salter A. J., (2013), "Cui Bono? The selective revealing of knowledge and its implications for innovative activity," *Academy of Management Review*, vol. 38, no. 2, pp. 270–291.

14. Asay C. D., (2016), "The information Value of Patents," *Berkeley Technology Law Journal*, vol. 31, no. 1, pp. 259–324.

15. Sundaresan, S., Jena, D., and Nerkar, A., (2017), "Open(ing) IP: The Effects of Patent Non-Assertion on Innovation,", working paper.

16. Contreras, J. L., Hall, B. H., and Helmers, C., (2018), "Assessing the Effectiveness of the Eco-Patent Commons: A Post-mortem Analysis," CIGI Papers, no. 161.

17. Chander A. and Sunder M., (2004), "The Romance of the Public Domain," *California Law Review*, vol. 92, no. 5, pp. 1331–1373.

18. Contreras J. L., (2015), "Patent Pledges: Between the Public Domain and Market Exclusivity," *Michigan State Law Review*, pp. 787–792.

19. Chesbrough H. W., (2003), "The logic of open innovation: managing intellectual property," *California Management Review*, vol. 45, no. 3, pp. 33–58.

20. Contreras J. L., (2015), "Patent Pledges," *Arizona State Law Journal*, vol. 47, no. 3, pp. 543–608.

21. Lerner J., Strojwas M., and Tirole J., (2007), "The design of patent pools: the determinants of licensing rules," *The RAND Journal of Economics*, vol. 38, no. 3, pp. 610–625.

22. Contreras, J. L., "Non-SDO Patent Statements and Commitments," American University Washington / College of Law, 2019. [Online]. http://www.pijip.org/non-sdo-patent-commitments/. [Accessed: 03-Feb-2019].

23. Awad, B., (2015), "Global Patent Pledges: A collaborative mechanism for climate change technology," CIGI Papers, vol. 81.

24. Hsieh H.-F. and Shannon S. E., (2005), "Three Approaches to Qualitative Content Analysis," *Qualitative Health Research*, vol. 15, no. 9, pp. 1277–1288.

25. Kondracki N. L., Wellman N. S., and Amundson D. R., (2002), "Content analysis: Review of methods and their applications in nutrition education," *Journal of Nutrition Education and Behavior*, vol. 34, no. 4, pp. 224–230.

26. Charmaz K., *Constructing grounded theory*: *A practical guide through qualitative analysis*, vol. 10. SAGE Publications, 2006.

27. Saldaña J., *The Coding Manual for Qualitative Researchers*. Thousand Oaks, CA: SAGE Publications, 2009.

28. Gioia D. A., Thomas J. B., Clark S. M., and Chittipeddi K., (1994), "Symbolism and Strategic Change in Academia: The Dynamics of Sensemaking and Influence," *Organization Science*, vol. 5, no. 3, pp. 363–383.

29. Corley K. G. and Gioia D., (2004), "Identity Ambiguity and Change in the Wake of a Corporate Spin-Off," *Administrative Science Quarterly*, vol. 49, no. 2, pp. 173–208.

30. Nickerson R. C., Varshney U., and Muntermann J., (2013), "A method for taxonomy development and its application in information systems," *European Journal of Information Systems*, vol. 22, no. 3, pp. 336–359.

31. Doty D. H. and Glick W. H., (1994), "Typologies as a Unique Form of Theory Building: Toward Improved Understanding and Modeling," *Academy of Management Review*, vol. 19, no. 2, pp. 230–251.

32. Cornelissen J., (2017), "Editor's Comments: Developing Propositions, a Process Model, or a Typology? Addressing the Challenges of Writing Theory Without a Boilerplate," *Academy of Management Review*, vol. 42, no. 1, pp. 1–9.

33. Siebrasse N. V. and Cotter T. F., "Judicially Determined FRAND Royalties," in *The Cambridge Handbook of Technical Standardization Law: Competition, Antitrust, and Patents*, Contreras J. L., Ed. Cambridge University Press, 2017, pp. 365–388.

34. Sidak G. J., (2013), "The meaning of frand, part I: Royalties," *Journal of Competition Law and Economics*, vol. 9, no. 4, pp. 931–1055.

35. Varian H. R., *Microeconomic Analysis*. Norton, 1992.

36. Wang T., Venkatesh R., and Chatterjee R., (2007), "Reservation Price as a Range: An Incentive-Compatible Measurement Approach," *Journal of Marketing Research*, vol. XLIV, May issue, pp. 200–213.

37. Miller K. M., Hofstetter R., Krohmer H., and Zhang Z. J., (2011), "How should consumers' willingness to pay be measured? An empirical Comparison of state-of-the-art approaches," *Journal of Marketing Research*, vol. XLVIII, February issue, pp. 172–184.

38. Buckland S. T., Macmillan D. C., Duff E. I., and Hanley N., (1999), "Estimating Mean Willingness to Pay from Dichotomous Choice Contingent Valuation," *Journal of the Royal Statistical Society*, vol. 48, no. 1, pp. 109–124.

39. Poe G. L., Welsh M. P., and Champ P. A., (1997), "Measuring the Difference in Mean Willingness to Pay When Dichotomous Choice Contingent Valuation Responses Are Not Independent," *Land Economics*, vol. 73, no. 2, pp. 255–267.

40. Poltorak A. I. and Lerner P. J., *Essentials of Intellectual Property Licensing*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2004.

41. ThePatentPledge.org, "The Patent Pledge," 2019. [Online]. www.thepatentpledge.org. [Accessed: 04-Feb-2019].

42. Tesla Motors, "Tesla Motors Patent Pledge," 2019. [Online]. www.tesla.com/en_GB/about/legal#patent-pledge. [Accessed: 04-Feb-2019].

43. Eisenhardt K. M., (1989), "Building Theories From Case Study Research," *Academy of Management Review*, vol. 14, no. 4, pp. 532–550.

44. Martin R. and Sunley P., (2003), "Deconstructing clusters: chaotic concept or policy panacea?," *Journal of Economic Geography*, vol. 3, pp. 5–35.

45. Teece D. J., (1986), "Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy," *Research Policy*, vol. 15, February issue, pp. 285–305.

46. Hanel P., (2006), "Intellectual property rights business management practices: A survey of the literature," *Technovation*, vol. 26, no. 8, pp. 895–931.

47. Al-Aali A. Y. and Teece D. J., (2013), "Towards the (Strategic) Management of Intellectual Property: Retrospective and Prospective," *California Management Review*, vol. 55, no. 4, pp. 15–30.

48. Resnik D. B., (2004), "Fair drug prices and the patent system," *Health Care Analysis*, vol. 12, no. 2, pp. 91–115.

49. Bogers, M., Bekkers, R., and Granstrand, O., (2012), "Intellectual Property and Licensing Strategies in Open Collaborative Innovation," forthcoming in: Open Innovation in Firms and Public Administrations, pp. 37–58.

50. Rockett K. E., (1990), "Choosing the Competition and Patent Licensing," *Journal of Economics*, vol. 21, no. 1, pp. 161–171.

51. Grindley P. C. and Teece D. J., (1997), "Managing intellectual capital: Licensing and Cross-Licensing in Semiconductors and Electronics," *California Management Review*, vol. 39, no. 2, pp. 8–41.

52. Shapiro C., (2001), "Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard-Setting," *Innovation Policy and the Economy*, vol. 1, pp. 119–150.

53. Arora A., Fosfuri A., and Gambardella A., (2001), "Markets for Technology and their Implications for Corporate Strategy," *Industrial and Corporate Change*, vol. 10, no. 2, pp. 419–451.

54. Eppinger E., *Patentpools—Eigenschaften, Motive und Implikationen*. Springer Fachmedien Wiesbaden, 2015.

55. Maher J. H., *Software Evangelism and the Rhetoric of Morality: Coding Justice in a Digital Democracy*. New York City: Taylor & Francis, 2016.

56. den Uijl S., Bekkers R., and de Vries H. J., (2013), "Managing Intellectual Property Using Patent Pools: Lessons from Three Generations of Pools in the Optical Disc Industry," *California Management Review*, vol. 55, no. 4, pp. 31–50.

57. LOT-Network, "Homepage," 2019. [Online]. www.lotnet.com. [Accessed: 22-Mar-2019].

58. Pohlmann T. and Opitz M., (2013), "Typology of the patent troll business," *R&D Management*, vol. 43, no. 2, pp. 103–120.

59. Bessen J. and Meurer M. J., *Patent Failure*. Princeton: Princeton University Press, 2009.

60. Diessel B. H., (2007), "Trolling for trolls: The pitfalls of the emerging market competition requirement for permanent injunctions in patent cases post-eBay," *Michigan Law Review*, vol. 106, no. 2, pp. 305–345.

61. Franzinger M. R., (2003), "Latent Dangers in a Patent Pool: The European Commission's Approval of the 3G Wireless Technology Licensing Agreements," *California Law Review*, vol. 91, pp. 1693–1728.

62. Fosfuri A., (2006), "The licensing dilemma: Understanding the determinants of the rate of technology licensing," *Strategic Management Journal*, vol. 27, no. 12, pp. 1141–1158.

63. Katz M. L. and Shapiro C., (1986), "Technology Adoption in the Presence of Network Externalities," *Journal of Political Economy*, vol. 94, no. 4, pp. 822–841.

64. da Silva M., (2017), "Open innovation and IPRs: Mutually incompatible or complementary institutions?," *Journal of Innovation & Knowledge*, in press.

6.11 Discussion Topics

The topic for discussion in this chapter is on the types of copyright, patent and trademark infringement in the digital world. The digital age has opened up a number of opportunities for sharing and accessing creative works, technological innovations, and brand identities. But with these advances, the instances of copyright, patent and trademark infringement have also increased. This forum provides a comprehensive discussion of the intellectual property challenges of the digital age. They cover legal, economic, technological and ethical aspects, thus offering a solid basis for discussions on the changing character of copyright, patent and trademark infringement in a rapidly digitizing world.

Here are some suggested discussion topics on copyright, patent and trademark infringement in the digital age:

1. How digital platforms (e.g. YouTube, Spotify, social media) handle copyright violations.
2. The consequences of user generated content on the spreading of copyrighted materials.
3. The legal duties of platform operators in preventing copyright infringement (e.g. YouTube's DMCA safe harbor).
4. The extent of online piracy in the music, film and gaming sectors.
5. The efficiency of anti-piracy measures (e.g. DMCA takedown notices, site blocking).
6. The effects of piracy on content creators and the entertainment industry.
7. The role of international cooperation in fighting piracy.
8. Penalties for businesses that breach open source licenses.
9. The legal and ethical aspects of the use of pirated software in the workplace.
10. How counterfeit goods are marketed and sold through online platforms like Amazon and eBay.
11. The role of social media influencers and digital marketing in trademark infringement.
12. The legal consequences of cybersquatting and domain name disputes.
13. The difficulties of enforcing trademarks in different jurisdictions in a globalized economy.
14. How fair use applies to digital content, such as memes, parodies, and commentary videos.
15. The growth of content creation on platforms such as TikTok and the potential for copyright infringement.
16. Recent legal cases and court decisions that have influenced the fair use doctrine in the digital era.
17. How the 'transformative' nature of content influences fair use claims in digital environments.
18. Who owns the data generated by users on digital platforms? The owners of the platform or the users?
19. Data scraping, aggregation, and ownership disputes: The challenge of striking a balance between the public interest and proprietary rights.
20. How illegal streaming services get around copyright laws and enforcement.
21. The economic effect of illegal streaming on the entertainment industry including film, sports and music.
22. How social media changes conventional views of copyright especially regarding short, viral content.

Make use of the knowledge you gain from this book and your own experience, opinion and judgment will make your post more attractive for discussion.

6.12 Projects

6.12.1 Copyright Contentions: The Legal and Ethical Battle Over Digital Copyright Management

Abstract: This project examines a legal dispute between a website operator and an author who sued the platform after it stored her work for its filter database while the site defended its actions. Throughout four weeks this research examines the legal arguments of both sides and reviews established case law while studying the ethical aspects of copyright management in digital platforms. Through this research project we will achieve two goals by studying copyright law as it applies to content hosting platforms while deciding this case through assessment of the presented evidence.

Research Questions: When a website stores copyrighted material for preventing unauthorized reposting does this action violate the author's copyright rights? What legal rights and duties does the website possess from its actions and how do these match existing copyright laws? Which moral principles affect the methods through which content owners protect their work against copyright infringement?

Methodology: Legal Research: Examine current copyright laws, relevant legal cases, and the doctrine of failuse. Argument Construction: Legal research and precedent will guide the creation of arguments supporting both the author and the site during this analysis. Mock Court Proceedings: A mock trial simulation will allow participants to hear and argue the case through legal frameworks and evidence. The analysis of arguments will lead to a verdict that makes a legal assessment of the case's strengths and weaknesses.

Expected Outcomes: The report will include thorough arguments from both sides along with essential copyright background information. The research includes a mock trial transcript that shows both legal and ethical aspects of the case. This project will present an opinion which considers both legal aspects and ethical issues of the case. Content hosting platforms should follow copyright management recommendations which balance protection for authors with operational content control capabilities.

Timeline:

- Week 1: Planning and Initial Research. An outline and project schedule will be established at this time. Start by researching copyright laws and deciding which cases to use as examples.
- Week 2: Argument Development. The website should present its legal objections against the author in this step. The defense will be established through safe harbor provisions together with fair use arguments presented by the site.
- Week 3: Mock Trial Preparation. The mock trial requires organization to determine which participants will serve as prosecution defense and judges. The team should create opening statements and evidence exhibits and final closing arguments for the trial.
- Week 4: Mock Trial Execution and Final Analysis. A thorough debate of the presented arguments will take place during the mock trial execution. The analysis will help create a verdict by examining the presented evidence together with the arguments. The final report will incorporate research findings and suggested recommendations.

Conclusion: Through this project students will understand copyright law complexities in digital environments as well as intellectual property management ethics and website practical considerations for user-generated content hosting.

6.12.2 The Ethical and Legal Debate on Digital Rights Management and Fair Use

Abstract: The study investigates the dispute between Digital Rights Management (DRM) systems and their possible violation of copyright users' fair use provisions during a four-week undergraduate research period. The research investigates both DRM implementation ethics from content creators' and consumers' perspectives regarding their ability to circumvent such systems for legal fair use purposes. The research presents an equitable perspective between creators' IP protection and users' rights to perform fair use according to legal frameworks.

Research Questions: The ethical and legal basis exists for content creators to implement DRM systems as intellectual property protection measures. What legal and moral boundaries exist for consumers who want to use circumvent DRM restrictions for fair use purposes? The application of current copyright laws and fair use doctrine along with their implications for DRM use and circumvention remains unclear.

Methodology: A systematic evaluation of DRM and fair use literature includes legal texts and academic publications together with relevant case studies.

Surveys: The project includes surveys that aim to capture the viewpoints of content creators alongside their target audience regarding DRM and fair use rights.

Legal Analysis: The study investigates both current legal statutes and established precedents that control DRM applications and fair use regulations throughout different nations.

The ethical theories will be used to evaluate the moral arguments that exist for and against DRM and its circumvention.

Expected Outcomes: The report will present a detailed analysis of both ethical and legal aspects regarding DRM implementation and user activities around it. The report will deliver an objective evaluation between copyright holders' rights and users' freedoms. A summary of survey findings revealing the perceptions and attitudes of stakeholders towards DRM. The research will present policy recommendations along with industry advice to resolve DRM versus fair use rights conflicts.

Timeline:

- Week 1: Project Planning and Literature Review. The project framework and objectives need to be defined first. The research focuses on DRM and fair use doctrine through an initial literature review during Week 1.
- Week 2: Surveys and Legal Research. The research team needs to create survey instruments before they can collect stakeholder opinions. The research team performs extensive legal research on existing laws and case precedents.
- Week 3: Analysis of Survey Data and Ethical Considerations. Survey response analysis will measure stakeholder sentiment about DRM and fair use practices. The data along with legal findings will serve as inputs for ethical reasoning to guide the discussion.
- Week 4: Synthesis and Final Reporting. The research team will integrate survey results with legal findings together with ethical arguments into a unified report. The research team will create a concluding section that unifies ethical and legal findings by drafting the final report. The research team will create an informative presentation to present the findings together with suggested recommendations.

Conclusion: The research project aims to develop understanding about IP protection versus public fair use rights which could guide future legal developments and industry regulations.

6.13 Exercises

1. Why did the rise of computer technology create challenges for intellectual property law?
2. What significant milestone was marked by the U.S. Patent Act of 1790?
3. What is the purpose of the Digital Millennium Copyright Act (DMCA) of 1998?
4. What are some key issues in balancing intellectual property protection and fostering innovation in computer technology?
5. How does the concept of copyright differ when applied to software compared to traditional creative works?
6. How has online piracy contributed to copyright infringement?
7. What is plagiarism in the context of copyright infringement?
8. Why can unauthorized remixes and mashups be considered copyright infringement?
9. What is patent infringement?
10. How does cybersquatting infringe on trademark rights?
11. Why is it challenging to protect intellectual property in the digital age?
12. What is the fair use doctrine in copyright law?
13. What is the purpose of the fair use doctrine?
14. What are the four factors used to determine fair use?
15. How does the purpose and character of the use affect fair use?
16. Why is the nature of the copyrighted work important in fair use analysis?
17. How does the amount and substantiality factor influence fair use decisions?
18. Give examples of activities often protected by fair use.
19. Why is fair use important in education?
20. What is a transformative use, and why does it support fair use?
21. How does fair use foster creativity and innovation?
22. What is the Digital Millennium Copyright Act (DMCA) of 1998?
23. What is the "safe harbor" provision in Title II of the DMCA?
24. How does the DMCA's anti-circumvention rule affect technology developers?
25. What is the purpose of the "notice-and-takedown" procedure in the DMCA?
26. Why is the DMCA's impact on fair use controversial?
27. How does the DMCA benefit content creators and rights holders?
28. What are some criticisms of the DMCA?
29. How do consumers view the DMCA's anti-circumvention provisions?

30. Why is government involvement crucial in protecting intellectual property (IP)?
31. How do copyright laws support creators?
32. What is the purpose of patent laws?
33. How do trademark laws benefit consumers?

Chapter

7

Evaluating and Controlling Technology

- The importance of ethical evaluation in the development of technology
- Strategies for assessing the impact of technology on society
- The Neo-Luddite Critique of the Digital Age
- The role of stakeholders in regulating technology
- The ethical implications of emerging technologies, such as AI and biotechnology

7.1 The Importance of Ethical Evaluation in the Development of Technology

Modern technology remains a core component of our daily existence as it transforms our methods of interaction communication work and education and relations with our environment. Modern technology continues to transform various sectors of society through innovations in artificial intelligence and biotechnology as well as autonomous vehicles and social media platforms. Great responsibility accompanies great power which requires technology developers to adopt ethical approaches when creating and deploying their systems. Ethical evaluation functions as a vital guidance system for technological development to achieve maximum benefits and reduced potential risks. This session examines how ethical evaluation shapes technology development through various complex processes [175, 215, 112, 143, 226].

Protecting Human Values: Technology holds the power to transform how humans interact with their experiences and economic systems and social frameworks. Technology needs to operate within human values and rights boundaries to prevent accidental side effects from occurring. Technology evaluation systems analyze how technological systems affect privacy rights and security measures and user autonomy and equality standards. The increasing use of data-driven technologies creates issues about individual privacy rights and surveillance practices. Developers receive guidance from ethical evaluation to develop methods which protect user data anonymity while upholding individual rights throughout the digital era.

Protecting against Bias and Discrimination: The increasing use of technology in decision-making processes makes the problem of bias more critical. The biases within training data used by algorithms and AI systems transfer to generate unfair discriminatory results. Developers must use ethical evaluation to detect potential biases in their technology while taking steps to reduce them. The implementation of technology requires designers to create systems which guarantee fairness through equal treatment for every individual in all application domains including hiring systems and criminal justice systems and lending algorithms.

Unintended consequences should be minimized through appropriate measures. Through ethical evaluation technology developers gain a comprehensive understanding of how their systems affect social systems and environmental systems and cultural systems. Developers can take preventive measures against risks when they forecast negative impacts. Social media experienced explosive growth which created new ways to communicate but introduced problems including false information spread and digital harassment and

psychological health problems. The evaluation process helps developers weigh advantages and disadvantages of such platforms.

Long-Term Sustainability: The pursuit of innovation must not harm the sustainability that extends beyond short-term goals. The evaluation process for technology requires analyzing its environmental effects so developers can create sustainable green solutions. The tech industry produces major ecological effects through its energy usage and electronic waste generation. The ethical examination of technological solutions leads to research into renewable power generation as well as efficient equipment development and sustainable waste management.

The development of transparent systems with accountable mechanisms builds trust among the public toward technology. Ethical evaluation requires developers to provide detailed information regarding technological operations as well as the impact these systems have on user experiences. Developers must establish accountability systems which enable them to answer for their creations to the public. The combination of openness and accountability enables prompt corrective measures after ethical lapses or mistakes occur thus promoting a culture of accountable innovation.

The development of technology should include all groups of people. Developers who participate in ethical evaluation make accessibility and inclusivity essential elements during their initial development process. Technology functions as a powerful empowerment tool for social progress when developers focus on creating interfaces that serve disabled users and expand internet access to disadvantaged areas.

In conclusion, technology development depends on ethical evaluation as its fundamental principle. The development of innovation receives direction from ethical standards which protects human values while blocking bias and reducing unintended results and ensuring sustainability and promoting transparency and inclusivity. The world's ongoing transformation through technology makes ethical aspects of development increasingly vital. The integration of ethics as a central component in technological innovation will lead to a future where technology functions as a positive force to enhance human welfare while protecting our mutual beliefs and aspirations.

7.2 Strategies for Assessing the Impact of Technology on Society

Modern technology has created a profound and extensive impact on society during this era of quick innovation. New technologies require careful assessment of their potential impacts to guarantee their alignment with societal requirements and ethical principles and strategic objectives. The evaluation process of technology's impact on society requires a detailed examination that includes both temporary and permanent effects. The discussion explores methods for proper technology impact assessment on society while explaining why thorough evaluation is needed for creating responsible technological progress.

1. Multidisciplinary Research: Technology assessment for social impact needs cooperation between engineering with social sciences together with ethics alongside economics and additional disciplines. Research from multiple disciplines creates a complete picture of technology's effects across various societal elements. AI-driven healthcare systems require analysis of both their practical implementation along with ethical aspects and access considerations and financial sustainability and doctor-patient relationship impacts.
2. Ethical Frameworks and Guidelines: Ethical frameworks serve as a systemized method for analyzing the moral aspects of technology. The development of ethical guidelines alongside their implementation enables the assessment of negative consequences together with prejudice and conflicts which emerge from technological solutions. When designing autonomous vehicles developers must establish methods through which the system makes ethical choices in complicated situations so both drivers and pedestrians remain safe while the technology minimizes damage.

3. Scenario Analysis and Modeling: Scenario building and modeling allow the examination of multiple potential future scenarios which emerge from technology advancement. This approach predicts the effects of multiple technological development paths and their societal implications. The use of scenario analysis helps determine possible advantages and disadvantages which emerge from different strategic choices. Through modeling systems researchers can predict how renewable energy systems would impact the environment relative to fossil fuel power generation.
4. Longitudinal Studies: Extended observation through longitudinal studies enables researchers to determine the lasting consequences of technological advancements. These studies help researchers understand technology development along with its consequences for societal structures. The results of longitudinal research reveal patterns alongside unforeseen results and the capacity of systems to adjust to technological transformations. Analyzing social media platforms throughout time demonstrates how communication methods and mental health effects alongside cultural developments change.
5. Public Engagement and Participatory Approaches: Stakeholder perspectives together with public opinions represent essential factors for assessing societal impact. The combination of public participation methods and engagement approaches helps organizations collect multiple perspectives about technology from diverse stakeholders. By involving everyone throughout the process the developers can guarantee that their technologies match social values and requirements. Urban planning technologies developed through community participation in local planning processes produce designs which promote both livable environments and inclusive spaces.
6. Pilot Programs and Controlled Experiments: Pilot programs along with controlled experiments serve as testing mechanisms for new technologies which allow researchers to evaluate their performance before large-scale implementation. This research method reveals potential obstacles in implementation and enables designers to improve their work while revealing unanticipated outcomes. By running a new educational technology in a small test area educators can evaluate its success rate and determine its necessary improvements.
7. Continuous Monitoring and Evaluation: Society undergoes continuous changes because of technology. Continuous evaluation practices enable the assessment of ongoing impact while allowing necessary adjustments to be made. The strategy allows organizations to stay active and respond quickly to new challenges that emerge. The impact of digital platforms on children's mental health development necessitates ongoing observation together with continuous adaptation of established guidelines.

In conclusion, technology impact assessment on society requires a sophisticated approach which combines research from various fields together with ethical principles and future prediction models and long-term observation and public participation and experimental testing and sustained evaluation practices. These assessment methods enable a complete understanding of technology's effects which helps direct its development towards responsible progress and societal wellness.

7.3 The Neo-Luddite Critique of the Digital Age

The modern Neo-Luddite perspective about computer technology functions as an updated version of the original Luddite movement from the early 19th century to criticize fast technological progress in daily existence. Neo-Luddites analyze three aspects of modern technology including its effects on human relationships alongside privacy erosion and personal freedom reduction.

The fundamental basis of Neo-Luddism consists of questioning the automatic acceptance of technology and demanding a new evaluation of the positive nature of technological development. Neo-Luddites propose

society to embrace technology through conscious integration methods which should always prioritize human well-being above technical convenience. Key concerns include:

Privacy and Surveillance: Neo-Luddites fear the digital technology enables extensive surveillance practices that invade individual privacy rights. The surveillance programs operated by tech companies alongside government monitoring activities create risks to civil liberties.

The modern Neo-Luddites follow the original Luddite approach by demonstrating against automation and artificial intelligence because these advancements threaten to displace workers in different industries which might generate unemployment and economic imbalance.

Excessive dependence on digital communication has led to diminishing face-to-face interactions which causes social relationship deterioration alongside growing mental health issues including loneliness and depression.

According to Neo-Luddites high-tech devices create two major environmental problems by using up rare earth minerals and generating electronic waste.

The fear exists that technological dependence will eliminate human skills while destroying self-sufficient abilities in society. The Neo-Luddite movement fights to preserve our capacity to survive without complex technological systems.

The fast-paced development of technology in biotechnology and AI domains creates moral dilemmas according to Neo-Luddites because these issues receive insufficient attention from society.

The concentrated power of a few massive technology companies remains a significant concern because these entities dominate both market sectors and media platforms and political systems.

Neo-Luddites emphasize that the digital divide represents a major issue because it creates a gap between those who possess computer and internet access and those who lack it thus deepening social and economic inequalities.

The Neo-Luddism movement presents concerns about modern technology yet this movement does not oppose technology in all its forms. The majority of followers support proper technological development together with its appropriate utilization. They endorse technological developments which provide sustainability and equality while improving human life quality instead of reducing it. The movement supports public-interest-based regulations and grassroots movements which teach people to assess their technological decisions more thoughtfully.

In conclusion, the neo-Luddite perspective brings important wisdom to society's discussion about technology because technological progress exists to create a world that is simultaneously sustainable and humane. These groups work to establish dialogue that examines both what technology can achieve and what its proper actions should be.

7.4 The Role of Stakeholders in Regulating Technology

The fast-changing world of technological development requires strict regulation of new technology. The integration of technology into all aspects of society requires regulatory mechanisms to maintain ethical standards and legal compliance and societal acceptability. Technology regulation receives vital input from stakeholders who embody different social interests. This discussion examines the multiple facets of stakeholder involvement in technology regulation while demonstrating their essential contribution to ethical technological advancement.

1. Ethical standards for technology development emerge through collaboration between representatives from academia, industry, civil society and government bodies. The ethical framework requires consideration of privacy concerns as well as security measures and bias reduction and transparency

and human rights protection. The stakeholders' joint participation in discussions and negotiations enables them to establish frameworks which guarantee technology maintains societal values while avoiding potential risks.

2. The gap between new technology advancements and regulatory frameworks requires ongoing modifications because technological progress moves ahead of regulatory frameworks. Regulatory agencies together with industry experts and consumer advocates join forces to achieve a proper equilibrium between supporting innovation and protecting public safety. New medical device development requires stakeholders to work together for safety standard creation and testing protocols that protect patients from harm.
3. Technology regulation requires diverse stakeholder perspectives for effective implementation. Stakeholders embody different perspectives that encompass the views of vulnerable groups alongside those of underrepresented communities and future societal members. The inclusion of diverse stakeholders leads to regulations which eliminate gaps while making technological solutions accessible to every societal group.
4. The monitoring process of technology-related activities depends on stakeholder involvement for ensuring regulatory compliance. The adherence of technological practices to established standards is monitored by regulatory agencies together with industry associations and consumer watchdogs. The evaluation process includes checking data privacy practices and environmental impact and product safety and additional factors.
5. The rapidly evolving technology sector creates new regulatory problems which stakeholders need to solve through timely responses. Stakeholders use their domain knowledge to help detect new challenges and evaluate their importance. Stakeholders take the lead in updating regulations to respond to changing contexts through their work on ethical artificial intelligence and environmental consequences of new technologies.
6. Through stakeholder involvement the public obtains opportunities to participate in technology regulatory decisions. Through their efforts stakeholders organize public dialogues while collecting public feedback to ensure that regulations fulfill societal requirements. Public participation in data privacy regulation decisions helps policymakers create policies that match citizen expectations and societal values.
7. The implementation of transparent technology regulation methods serves as a foundation to establish public trust. Through their advocacy work stakeholders push for transparent decision-making processes that disclose all relevant technology development information and usage data along with identified potential risks. Through transparent practices the system reduces harmful outside control while increasing responsibility and uses verified information to develop regulations.
8. The globalized technological environment requires international stakeholders to work together for effective regulation. International organizations and governments together with industry bodies work with stakeholders to create standardized regulations and standards. The collaborative work enables technology to follow identical principles across all geographical regions.

In conclusion, stakeholders play an essential and diverse role in the regulation of technology. The comprehensive and responsible regulatory framework emerges through stakeholder involvement in defining ethical standards and maintaining innovation-safety equilibrium and representing diverse perspectives and enforcing regulations and addressing emerging challenges and promoting public engagement and advocating for transparency and facilitating international collaboration. Through their active participation technology advances social progress while protecting cultural values and reducing threats and delivering maximum advantages to present and future populations.

7.5 The Ethical Implications of Emerging Technologies, Such as AI and Biotechnology

Emerging technologies including Artificial Intelligence (AI) and biotechnology have been introduced in the 21st century because of the quick technological development. Along with these opportunities there are ethical concerns which need examination. This session discusses the ethical issues related to two revolutionary technologies: Artificial Intelligence (AI) and biotechnology, and the problems that arise for society.

Artificial Intelligence (AI):

1. The learning process of AI systems uses extensive data which occasionally brings forward the bias found in the information. The discriminatory outcomes from biased AI decisions affect various domains including employment selection and financial services and criminal justice. The ethical issue arises from the need to develop training systems for AI which utilize diverse datasets and from creating bias-detection tools to ensure fairness.
2. Self-operational AI systems including drones and autonomous cars generate doubts about accident responsibility. The problem of assigning liability to machines during their decision-making processes with major life-changing outcomes conflicts with established liability systems. The establishment of ethical frameworks needs to resolve issues regarding responsibility and accountability for these situations.
3. The capabilities of AI-powered surveillance systems to breach individual privacy create substantial issues about the collection of personal information and personal freedom. The ethical challenge to maintain public safety together with individual privacy demands strong regulatory measures alongside open disclosure of AI surveillance system operations.
4. Widespread AI and automation implementation threatens specific job sectors leading to both job displacement and increased economic inequality between individuals. The ethical necessity for societal protection requires policymakers to establish just transition programs for workers along with training and skill development initiatives to avoid societal damage.

Biotechnology:

1. Modern advances in biotechnology allow genetic modification of humans which creates new ethical issues about modifying human characteristics. The quest for genetic enhancement creates social distinctions between genetic levels which opposes basic equality principles and human dignity standards.
2. The capability to modify genes opens the possibility of designer babies allowing parents to choose particular characteristics for their offspring. The process of genetic enhancement raises ethical concerns about eugenics along with potential dangers for modifying the genetic composition of future offspring.
3. Research and commercial applications of genetic data raise ethical issues about informed consent and data protection and privacy. People must maintain control over the usage of their genetic information to protect their autonomy from exploitation.
4. The deployment of biotechnology can create environmental damage to ecosystems and biodiversity by releasing genetically modified organisms into natural environments. The lack of proper biosecurity measures and insufficient risk assessments could lead to unintended ecological consequences.

In conclusion, the ethical consequences of emerging technologies AI and biotechnology demand responsible actions during their development and deployment stages and their subsequent regulation. The solution to these ethical dilemmas demands joint work between developers of technology and ethicists and policymakers and wider society members. These technologies need proper ethical management because they will continue to advance our world while we must find a way to merge innovation with core values for beneficial contributions to human well-being.

7.6 Case Study 1

RESEARCH ARTICLE

Analyzing technology acceptance and perception of privacy in ambient assisted living for using sensor-based technologies

Wiktoria Wilkowska [1☯*], **Julia Offermann** [1☯], **Susanna Spinsante** [2☯], **Angelica Poli** [2‡], **Martina Ziefle** [1‡]

1 Human-Computer Interaction Center, RWTH Aachen University, Aachen, Germany, 2 Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy

☯ *These authors contributed equally to this work.*

‡ *AP and MZ also contributed equally to this work.*

* *wilkowska@comm.rwth-aachen.de*

Abstract

People increasingly use various technologies that enable them to ease their everyday lives in different areas. Not only wearable devices are gaining ground, but also sensor-based ambient devices and systems are increasingly perceived as beneficial in supporting users. Especially older and/or frail persons can benefit from the so-called lifelogging technologies assisting the users in different activities and supporting their mobility and autonomy. This paper empirically investigates users' technology acceptance and privacy perceptions related to sensor-based applications implemented in private environments (i.e., passive infrared sensors for presence detection, humidity and temperature sensors for ambient monitoring, magnetic sensors for user-furniture interaction). For this purpose, we designed an online survey entitled "Acceptance and privacy perceptions of sensor-based lifelogging technologies" and collected data from $N = 312$ German adults. In terms of user acceptance, statistical analyses revealed that participants strongly agree on the benefits of such sensor-based ambient technologies, also perceiving these as useful and easy to use. Nevertheless, their intention to use the sensor-based applications was still rather limited. The evaluation of privacy perceptions showed that participants highly value their privacy and hence require a high degree of protection for their personal data. The potential users assessed the collection of data especially in the most intimate spaces of domestic environments, such as bathrooms and bedrooms, as critical. On the other hand, participants were also willing to provide complete data transparency in case of an acute risk to their health. Our results suggest that users' perceptions of personal privacy largely affect the acceptance and successful adoption of sensor-based lifelogging in home environments.

Introduction

The current development in information and communication technologies enables a variety of applications in many different contexts. Assisting technologies such as wearable devices as well as lifelogging applications [1] capable of assisting users in their (smart) home environments or in Ambient Assisted Living (AAL) institutions allow for meaningful support in different areas of life [2, 3]. Especially within the context of healthcare and social care, lifelogging technology opens up far-reaching opportunities for preventive medicine,

https://doi.org/10.1371/journal.pone.0269642 Used with CC 4.0 license

diagnostics, therapeutical/rehabilitation measures and allows controlling healthcare cost. Particularly for older users, lifelogging technologies support autonomous living for longer time periods.

Modern technologies reach not only young and tech-savvy people who know how to use the almost daily newly emerging applications in their private and professional lives, but also those who did not directly grow up with the technology, having rather acquired its use at some later point in life. Especially older and/or frail persons can significantly benefit from the potential offered by assistive technologies in their private settings by increasing their health-related safety and maintaining their physical well-being and mobility.

Besides general applications that allow monitoring of physical activities, daily habits [4], vital parameters [5], or the detection of emergencies (e.g., [6]), older and/or frail persons can also use age-specific or illness-specialized functions. For instance, sensor-based technologies enable identifying changes in behavior, movement and sleeping patterns, or walking speed [7]; they also provide potential indicators for the recognition of diseases, such as dementia or Parkinson's disease (e.g., [8, 9]). Hence, analyses of such data enable an early identification of deviations from the normal behavior, health changes, or the detection of possible abnormalities in bodily functions as well as provide security for the users and their important others in the long run.

Despite the availability of different assistive applications, however, the main questions arising in this context are whether individuals are prepared for adopting such assistive technologies, to what extent they intend to use them, and where the possible barriers may lie. On the other hand, the lifelogging of (health-related) data from wearable or ambient technologies significantly interferes with the users' privacy and experience of autonomy. Research has shown that invasion of privacy poses one of the main concerns in the adoption of technologies (e.g., for voice-based digital assistants: [10]) as the assistive technologies capture a great amount of private information that people may not want to share with others [11].

Considering the fast pace of emerging assistive applications or systems, the current stage of research is insufficient when it comes to the ethical, legal, and social implications of using sensor-based technologies in home environments. There is a lack of a context-sensitive analysis of user acceptance and an empirically based exploration of the users' requirements for the recording and managing of (health-related) user data referring to their privacy. In this paper, we therefore address the current gap in research and empirically investigate the users' attitudes towards acceptance and privacy in the context of sensor-based technologies in private environments.

Theoretical background

In the following, we firstly describe the technology under study in more detail, and we highlight the theoretical background that provides the basic framework for the empirical investigation of user acceptance and perceptions of privacy for the adoption of sensor-based lifelogging technologies from the user's perspective.

Lifelogging technologies

As we focus in this study on lifelogging technologies, in the first step, we briefly outline the definition of the term. Selke [1] described "lifelogging" as different types of digital self-tracking and recording of everyday life and a phenomenon that can be placed between innovative technologies and cultural transformation. Lifelogging refers to diverse types of self-tracking, ranging from health prevention and the monitoring of bodily functions over one's location and motion to the measurement of productivity at work and in private life. Thus, the technology can capture and record certain parts of a human life in real-time, saving them for later use and analyses.

https://doi.org/10.1371/journal.pone.0269642

With this in mind, lifelogging technologies have a great potential to manage many of the concerns raised by population aging. The applications can be used to prevent disease, provide personalized healthcare, and give support to formal and informal caregivers [12]. There are different types of technologies, which can be applied for the mentioned purposes: Most common are wearable devices, such as smart watches and wristbands, often used for the monitoring of physical activities, location determination, and general self-awareness. This trend of using lifelogging for self-improvement is predominantly associated with younger and physically fit adults [13, 14] and it gains particular interest when it comes to the topics of health prevention and fitness. However, lifelogging systems can also provide great support for older and/or frail people by integrating them into the everyday life of persons in need of help. This means that lifelogging applications can not only be worn on the body, but they can be also integrated into a domestic space in order to capture useful data; in this context, the technology is then classified as AAL technology [15], shifting the focus to the function of assistance rather than self-improvement. Information gained in this way can be monitored by caring relatives and/or nursing personnel, and the data can be further analyzed by professionals to identify irregularities or emergencies, or even to predict changes in behavior or health status [16].

Generally, the process of lifelogging is considered a passive and non-intrusive activity which can be carried out by various types of lifelogging systems [13], ranging from simple trackers over (wearable) cameras and microphones to ambient sensors integrated into living environments. Accordingly, the recorded data can consist of multimedia data like video and audio recordings as well as sensor data [17].

Ambient sensor-based technologies used in the study

As mentioned above, despite being mostly referred to the domain of wearable sensors and devices, lifelogging may also rely on sensing infrastructures or sensor systems displaced in living environments. Setting up an ambient sensor-based lifelogging solution requires the installation of hardware devices in the living environment and their connection to a gateway that acts as a data aggregator and relay to a remote server. Typically, on the remote server, the received data is organized and presented to the user by a dashboard, from which several functionalities of reporting, analytics, and visualization are accessible.

Focusing on the selection of ambient sensors and their installation, a basic lifelogging setup, like the one considered in this study, may be obtained by using a small set of simple sensors, such as Passive InfraRed sensors (known as PIR)—to capture motion and presence in a given environment or room—, relative humidity and temperature sensors (H&T), and magnetic sensors (MAG) to detect doors and windows opening and closing but also user's interaction with drawers or furniture, depending on the specific installation location. Both PIR and MAG sensors may be classified as binary sensors, as their output can reflect two possible states only, such as presence/no presence or motion/no motion, and open/close, respectively. H&T sensors, on the contrary, provide numerical information that can be used not only to monitor indoor comfort, but also to detect specific events in the living environments, correlated to the user's habits. All these types of sensors are easily found in the market and sold by several different vendors, with prices starting from a few Euros to several tens Euros per device, depending on the brand chosen, the desired accuracy and precision of the sensing capability, the availability of additional services supported by the vendor, usually through a mobile app and a dedicated server for which the user has to pay a monthly or annual access fee. For a quick and minimally intrusive installation, nowadays almost all the sensors are battery-powered and enabled with a Bluetooth Low Energy (BLE) or ZigBee wireless connection to a specific device, named gateway (or concentrator), in charge of collecting the data generated by many sensors and relay it to the remote server, usually over a WiFi link.

In our study, the PIR sensor was located inside the bathroom, on the top of the mirror, in order to identify the events corresponding to the user entering and exiting the bathroom. According to the specifications provided by the manufacturer, the detectable distance amounts to 7 m, with a 170° field of view. The chosen H&T sensor can detect a 0.3°C temperature variation in the range [0, 60] °C, and a 3% variation of the relative

humidity in the range [0, 99] %. The sensor was located inside the kitchen to detect moments of the day related to food preparation (i.e., cooking events). MAG sensors have been applied to drawers and furniture doors to collect data able to describe the way people interact with the living environment: specifically, one MAG sensor was applied onto the fridge door, another one onto the drawer containing silverware. All the sensors were bought together, in a kit including the gateway as well, for around 80 Euros. An associated free-to-download Android mobile app is available, to set up and configure the sensors, and to visualize the list of collected data, but no additional services or remote dashboard are provided.

Technology acceptance of sensor-based lifelogging technologies

In the past, many scientific studies have shown that the adoption of a relevant technology is not solely determined by its technical merit. User acceptance is almost equally decisive for the successful establishment. However, the technology acceptance is a complex construct, and its nature may differ depending on the context under study.

In many years of research on technology acceptance meaningful models have been developed. To date, the Technology Acceptance Model (TAM, [18]) and the Unified Theory of Acceptance and Use of Technology (UTAUT; [19, 20]) have become established in this area. In this paper we focus on the TAM as this model has built the theoretical background for the current study. TAM is founded upon the hypothesis that the technology's acceptance and use can be explained in terms of the user's internal beliefs, attitudes, and intentions [21]. The model assumes that an individual's behavioral intention to use a technology or an information system is determined by the perceived ease of use (PEU) and perceived usefulness (PU). Davis [18] defined PEU as "the extent to which a person believes that using the system would be free of effort" (p. 320), claiming that an application perceived as easy to use is more likely to be accepted by the user. Further, the author defined PU as "the extent to which a person believes that using the system would enhance his or her job performance" ([18], p. 320) and assumes that technology or application that is perceived as highly useful leads to a positive use-performance relationship. With this in mind, technology considered useful is the one of which users believe that it helps them to perform the job, planned function, or action better. The model theorizes furthermore that PU is connected with PEU, because the easier the technology's use is, the more useful it can be perceived [22]. In addition, TAM claims that external variables indirectly influence technology use through PEU and PU, which affect the user's attitude (A) towards, and the behavioral intention to use (ItU), the technology. The theoretical model is schematically presented in Fig 7.1.

Even though the mentioned theoretical models have found very wide deployment in the field of information and communication technology, mainly addressing the job-related context of using, they do not satisfactorily

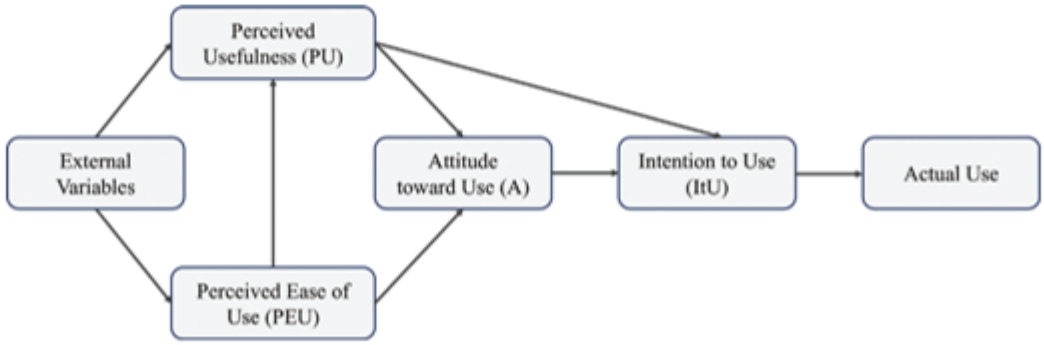


Fig 7.1 Technology acceptance model [18].
https://doi.org/10.1371/journal.pone.0269642.g001

evaluate all kinds of the currently available technologies. Especially in the area of the thriving lifelogging technologies, more specific opinions of the potential users are needed for the evaluation of the intention to use, and the real use, of the wearable or ambient applications. Previous research showed that while the perception of benefits can essentially encourage the utilization of a targeted technology, the awareness of barriers can significantly diminish the intention to use it (e.g., [23, 24]). Lifelogging applications which have the potential to assist the users—especially the older and/or frail ones—in the everyday life are mostly appreciated for the ability to support autonomy (e.g., [25]), the improvement of the sense of safety and security (e.g., [26, 27]), and easing the burden of caring family members [23, 28, 29]. Continuing within the context of aging in place, use of lifelogging applications can also be associated with significant concerns. Invasion of individual privacy, stigma, and feelings of isolation are important barriers in terms of using assistive lifelogging in the daily routine (e.g., [12, 30, 31]). As to data security, especially the perceived loss of control over the sensitive data, an unauthorized access to, or misuse of, that data by third parties represent great barriers to the use of lifelogging applications [11, 32].

Besides the perceived motives and barriers in the process of the adoption of assistive technologies, the properties of the technology itself as well as the characteristics of the users can have a decisive impact on the acceptance. Himmel and Ziefle [33] provided evidence that cameras were less desired as assistive lifelogging technology in comparison to motion sensors or microphones. The preferences for certain modes of technology were also validated in a study of emergency detection that showed wearable devices, such as smart watches or emergency buttons, being clearly more accepted by the users than cameras [34]. Notwithstanding, a study of Gelonch et al. [35] showed that older adults diagnosed with mild cognitive impairment can become competent users of lifelogging wearable cameras with a good level of acceptance. Hence, the adoption and actual use of different assistive applications clearly depend on the specific functionality and the circumstances of use, on the one side, and the characteristics of the receivers, on the other side.

Requirements of personal privacy

Perceptions of individual privacy pose a specific concern in the adoption of lifelogging technologies. As these technologies enable the user to capture and publish his/her personal data with ease, they also pose a major challenge in terms of maintaining the user's privacy.

The right to privacy is one of the fundamental human rights in a modern society and in the European Union it is protected by the General Data Protection Regulation (GDPR). According to Goddard [36], the GDPR has six core privacy principles that concern the general data protection: (1) fairness and lawfulness, (2) purpose limitation, (3) data minimization, (4) accuracy, (5) storage limitation, and (6) integrity and confidentiality. However, data protection by design and default is at the core of the GDPR and it is supported by transparency (full information is provided to individuals) and accountability (organizations take demonstrable responsibility using personal data). Mihailidis and Colonna [12] proposed the application of Privacy by Design (PbD) as one way to manage privacy concerns raised by lifelogging technologies. This approach involves the transformation of legal rules, namely those that pertain to privacy and data protection, into information systems and allows for incorporating privacy into all stages of a lifelogging system's development, considering the contextual understanding and the specific design elements of these kinds of systems.

Depending on the type of the lifelogging application, differently large amounts of user data are collected and stored. Research has shown that the expectation of privacy for the lifelogger can differ, depending on the context of use and the type of the used technology. As already mentioned above, video-based lifelogging poses a major hurdle to the acceptance and use of a technology [33, 34]. In this field, Padilla-López et al. [37] proposed a visual privacy by context approach that strives for a trade-off between the privacy preservation and the intelligibility of images being acquired. With this approach, the privacy protection can be constituted by different visualization modes, which conceal the sensitive information to a variable extent, providing thus

context-dependent levels of protection of the lifelogger [38]. According to different studies on the acceptance of assistive technologies, privacy concerns, especially those regarding the sense of surveillance and misuse of data access, are confirmed to be a barrier (e.g., [39]). The trade-off between the benefits provided by assistive technology (e.g., increased security and preservation of autonomy) and the loss of privacy is of major interest in acceptance research as it provides valuable insights into the (potential) users' decision processes towards the intention to use a particular technology. Research of Jaschinski et al. [40] revealed for instance that for older users, who are concerned about their privacy when using assistive technology, the benefits of increased safety and staying in their own home are significantly more relevant, and thus outweigh the feeling of the privacy loss. At this point, it becomes even clearer how sensible legal regulations are according to the PbD approach.

In the context of the present study, we investigate the perceptions of privacy with respect to ambient sensor-based lifelogging technologies considering not only different installation sites of the technology but also the different circumstances of the users. We additionally review how, in this context, individual privacy is related to the acceptance of this technology.

Objectives and aim of the study

On the basis of the presented considerations, previous research on the acceptance of assisting lifelogging technologies predominantly focused on video- and audio-based systems, and comparably few studies investigated the adoption of sensor-based technologies. In this study, we empirically investigate the users' attitudes towards acceptance and their perceptions of personal privacy in the context of sensor-based lifelogging in private environments. The underlying research questions of the study are the following:

RQ1: How high is the general acceptance of sensor-based technology in home environments: Do the potential users intend to use it? How do they perceive the technology's usefulness and ease of use?

RQ2: Are potential users more likely to be guided by perceptions of the benefits rather than the drawbacks that the technology brings?

RQ3: Which role does privacy play in the adoption of sensor-based ambient technologies and do the privacy perceptions change depending on the user's health condition?

RQ4: Does the protection of individual privacy have a significant impact on the acceptance and the willingness to use sensor-based technology?

Methods

In this section, we describe the methodological approach of our empirical study, introducing the structure of the online survey, the used study design, description of the sample, and the applied data analysis procedures.

Data collection

The study involving human participants was reviewed and approved by The Ethics Committee (Division 7.3) "Empirical Human Sciences" at the Faculty of Humanities at RWTH Aachen University (ID: 2021_003_FB7_RWTH Aachen). Prior to the survey beginning, participants were informed about the study's aim and content and provided their written consent to the scientific handling of their indicated data and to the participation in the study. Data collection was realized by means of a standardized online survey, which was performed between May and June 2021, using the professional platform Qualtrics XM.

The online survey was divided into three parts. The first part referred to the participants' demographic characteristics, i.e., age, gender, educational level, civil status, and housing situation. After this, participants described their own health condition [i.e., a general health status (ranging from 'very good' to 'very poor')],

https://doi.org/10.1371/journal.pone.0269642

chronic disease ('yes'/'no'; if 'yes': the need for support in everyday life), and their experiences of caregiving (professional vs. private experience in care).

The second part of the survey provided general information about sensor-based lifelogging technologies and systems. Here, the wide application spectrum of sensor-based lifelogging solutions was presented, pointing out both the common wearable and home-integrated sensors. As the focus of the study was to evaluate the acceptance of ambient sensors that assist (older/frail) inhabitants in their everyday life, concise information—previously elaborated by an interdisciplinary collaboration between engineering and communication science experts—was presented to the participants using descriptions and illustrations as presented in Fig 7.2. Therefore, the technical feasibility and the simple operability (on/off-functions) were exposed on the one hand, and all the potentials resulting from the possible overviews of daily habits, activities, and behaviors, as well as long-term comparisons, and useful profiles of different users, on the other. After receiving all the general details, participants were instructed to empathize with the situation, in which the introduced sensors are installed within their own home environment. Subsequently, they assessed the potential benefits (e.g., "ability to live independently at home") and barriers (e.g., "concerns about false alarms") of such lifelogging technology in their domestic environments. In addition, the evaluation of technology acceptance was extended by questions referring to criteria postulated by TAM [18]: perceived ease of use, perceived usefulness, and intention to use (6 items; Cronbach's $\alpha = .76$). All acceptance items were assessed on six-point Likert type scales ranging from "fully disagree" (= 1) to "fully agree" (= 6). Before starting with the next part, the survey evaluated the relevance of health in the perception of privacy when using sensor-based technology (the four items are presented in Fig 7.9; $\alpha = .80$), using the same (dis)agreement scale as for the acceptance statements.

Finally, the third part of the survey concentrated on the users' perceptions of privacy in the context of using sensor-based technology in home environments. First, participants rated their general perceptions of privacy in this context (e.g., "Privacy protection should be the first priority in sensor-based lifelogging.") using agreement scales as described above. Subsequently, the questionnaire collected opinions on privacy aspects connected to different specific activities, such as recordings of the stay in different premises, timing,

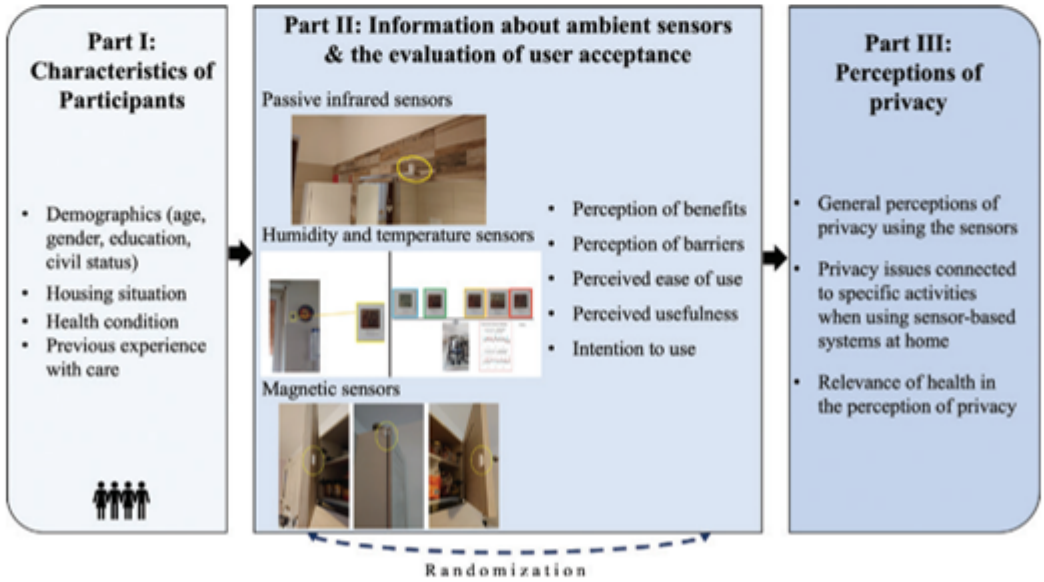


Fig 7.2 Schematic structure of the online survey.

https://doi.org/10.1371/journal.pone.0269642.g002

frequency and duration of individual activities, as well as generating of (temporal) profiles. Participants used a slider to rate how critical—in their eyes—such recordings of the respective activities are (from 0 ="not critical at all" to 100 ="very critical"). Fig 7.8 summarizes all the activities.

To ensure a high comprehensibility of the technical explanations and enable unambiguous projections of the technology object(s) to be evaluated, the survey was thoroughly pretested prior to the launch of the study. The respondents participated on a voluntary basis in this survey. Only adults were allowed to participate in the survey after they consented to a statement regarding the data protection and data retention.

Study design

The intention of this research was to examine the willingness to adopt sensor-based technologies in the private living environments. In the survey, we provide respondents with concrete information about the technology, and investigate to what extent such assistive technologies are accepted and permitted, or what conditions (e.g., privacy, data protection) need to be met in order to increase the acceptance among potential users.

Considering this, for the present study two main aspects are in focus: Firstly, the technology acceptance including the general perceptions of benefits (6 items; $\alpha = .90$) and barriers (6 items; $\alpha = .83$) associated with the use of sensor-based ambient technologies preferences as well as aspects like the perceived usefulness, ease of use, and the intention to use such sensors, applications, or systems; the later are described in detail in Table 7.1, while benefits and barriers are summarized in the Figs 7.4 and 7.5.

Secondly, the general perceptions of privacy as well as privacy in dependence on health pose another main aspect in the context of the sensor-based systems in this study; the used items are provided in Figs 7.7 and 7.9. To deepen the concept of privacy, the privacy assessments of individual activities when using ambient sensors complement the analyses: The respondents' evaluations of how critical they perceive the specific events/activities (see Fig 7.8) enable even more insights on privacy in the context of sensor-based lifelogging.

Both the mentioned aspects—technology acceptance and privacy perceptions—are assumed to represent the core premises of a successful adoption of the sensor-based technology, and are examined in this research in more detail. The study design is depicted in Fig 7.3.

Table 7.1 Items used for technology acceptance [18] adapted to sensor-based lifelogging technology and the internal consistency (Cronbach's α) of the particular criteria.

	Items	Reliability
Perceived usefulness (PU)	"It is useful to get an overview of the own activities with the help of sensor-based technologies." "With the help of sensor-based technologies, parts of life can be optimized."	$\alpha = .63$
Perceived ease of use (PEU)	"The use of sensor-based technologies seems very simple to me." "The use of sensor-based technologies is too complicated." [recoded]	$\alpha = .76$
Intention to use (ItU)	"I would not use a sensor-based lifelogging system for support in my home." [recoded] "I intend to use a sensor-based lifelogging system to assist me in my home in the future."	$\alpha = .75$
	Overall	$\alpha = .75$

https://doi.org/10.1371/journal.pone.0269642.t001

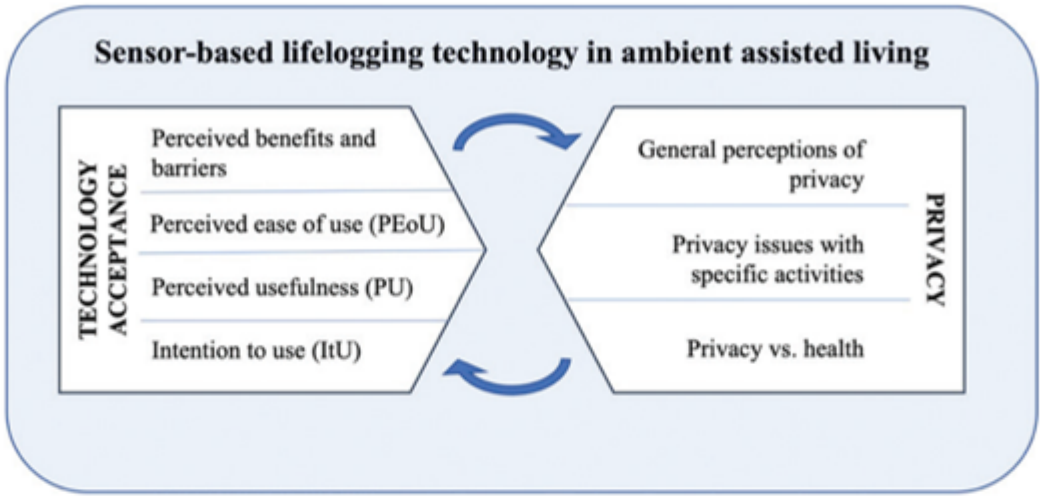


Fig 7.3 Study design.

https://doi.org/10.1371/journal.pone.0269642.g003

Sample description

After a quality inspection and data cleansing of our quantitative study, data from a total of N = 312 individuals were finally taken for the statistical analyses. The sample was rather young (M = 32.9 years; SD = 18.7) with participants ranging in their age between 18 and 91 years. There was a higher proportion of females (62.8%; n = 196) compared to males (36.9%; n = 115) in the sample, while one participant indicated a diverse gender (0.3%). On average, the participants were comparatively highly educated with 55.8% (n = 174) of them holding a university entrance qualification and 28.2% (n = 88) holding a university degree. Only 16.0% of the participants (n = 50) reported lower educational levels, such as secondary or elementary school certificates. As to civil status, most participants reported to live in a partnership or be married (56.4%; n = 176), one third (33.3%; n = 104) indicated to be single, and small parts of the sample were divorced (1.9%; n = 6) or widowed (4.2%; n = 13); the rest (4.2%, n = 13) did not indicate their civil status.

As the research focuses on the use of ambient sensor-based technologies, which, due to their multiple applications and support possibilities, can be beneficial especially for ill and/or elderly users, the health status of the participants was also of interest. The vast majority of the participants indicated a (very) good health condition (83.9%; n = 262). One fifth of all participants (20.2%, n = 63) reported to suffer from a chronic disease or have a physical disability or limitation. Accordingly, solely six respondents (1.9%) indicated to need assistance and support in their everyday life. Asked for experiences in care, 17.3% of the sample (n = 54) reported a professional experience in care, while 24% (n = 75) of them indicated an active private experience and 31.7% (n = 99) a passive private experience in caring for a person/family member in need of care. Regarding previous experiences in using lifelogging applications, respondents reported to use lifelogging for activity and health monitoring (e.g., number of steps, heart rate, sleeping pattern; 64.4%, n = 201), position tracking via GPS signal (66.7%, n = 208), the archiving of life events (i.e., collecting pictures, documents, or other data; 55.8%, n = 174), the tracking of consumption/health behavior (e.g., loyalty programs; 48.4%, n = 151), and performance measurements at work (9.3%, n = 29).

Data analysis

We perform descriptive analyses of the perceived benefits and barriers as well as TAM-related acceptance criteria and privacy issues related to the use of ambient sensor-based lifelogging technologies using means (M)

and standard deviations (SD); standard errors are provided in the graphs. Checking the internal consistency of the scales by means of Cronbach's Alpha ($\alpha > .7$) ensured a satisfying quality of the measured constructs (i.e., technology acceptance according to TAM [18], general privacy perceptions). To find out how privacy is predicting the technology acceptance, we calculated stepwise regression analyses. We set the level of statistical significance (p) at the conventional level of 5%.

Results

In the following, we present participants' attitudes towards using sensor-based technology in home environments in terms of technology acceptance and their perceptions of personal privacy. In the first part, we display the evaluations of perceived benefits and barriers, and we analyze the sensor-related technology acceptance in line with TAM [18]. In the second part of the Results section, we depict considerations as to privacy and its role for the use of ambient sensor-based technologies and applications. Within the third part, we estimate the relationships between the main research variables of this study.

Evaluations of user acceptance

Perceived benefits and barriers. Thinking of benefits when using different sensors at home, participants especially appreciate the function of possible notifications in case of emergency. The resulting means (Fig 7.4) indicate that items describing the emergency reach the highest values on the scale (>5 of 6 possible points; responses < 3.5 were considered disagreement and values > 3.5 were considered agreement). The individual sense of security ($M = 4.55$, $SD = 1.18$) and safety feeling for the relatives ($M = 4.78$, $SD = 1.17$) are also clearly appreciated. In addition, the ability to live independently ($M = 4.57$, $SD = 1.18$) and an untroubled use of the technology at home ($M = 4.44$, $SD = 1.27$) are perceived as beneficial, contributing to an accepted use of ambient sensors in home environments.

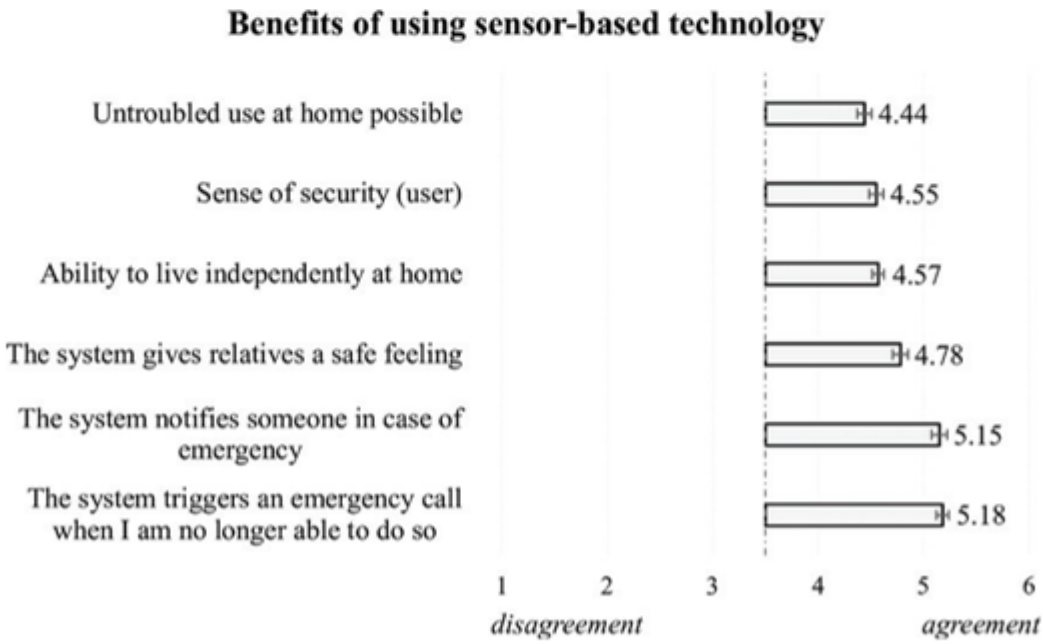


Fig 7.4 Perceptions of benefits when using sensor-based technology.

https://doi.org/10.1371/journal.pone.0269642.g004

Looking at findings for perceived barriers (Fig 7.5), it is noticeable that little agreement resulted for the studied items (all means are below the value of 3.5). According to the analysis, participants felt mostly uncertain with respect to the false alarms (M = 3.48, SD = 1.31), but they would use the technology not only in the event of a health-related restriction (M = 2.68, SD = 1.39). The respondents' positive attitude towards the sensors is also confirmed by the still quite low means resulting in the assessment of being monitored (M = 3.12, SD = 1.52) and regarding the potential fears of too high costs (M = 3.14, SD = 1.31). Participants disagreed that they would be disturbed by sensors installed in different rooms (M = 3.32, SD = 1.45) and their privacy concerns were also limited at this point (M = 3.31, SD = 1.45), although these aspects could potentially have a negative impact on the respondents' acceptance.

The positive perceptions of benefits and at the same time disagreements to possible barriers present a coherent picture, mirroring overall positive attitudes towards the ambient use of sensor-based technology.

Technology acceptance according to TAM. In order to underpin the aspect of acceptance in a more theoretically sound manner, the crucial technology acceptance criteria according to TAM [18] were examined.

The resulting means are depicted in Fig 7.6: While the mean values for the perceived ease of use (M = 8.75, SD = 1.97) and the ease of use (M = 8.79, SD = 2) clearly lie within the range of agreement (values >6.5), the intention to use ambient sensor-based lifelogging is rather restrained (M = 6.55, SD = 2.68), yet not rejected. In concrete terms, this means that the (potential) users perceive the ambient sensors as useful and predominantly easy to use, but they represent a rather neutral attitude towards the actual use of this technology so far.

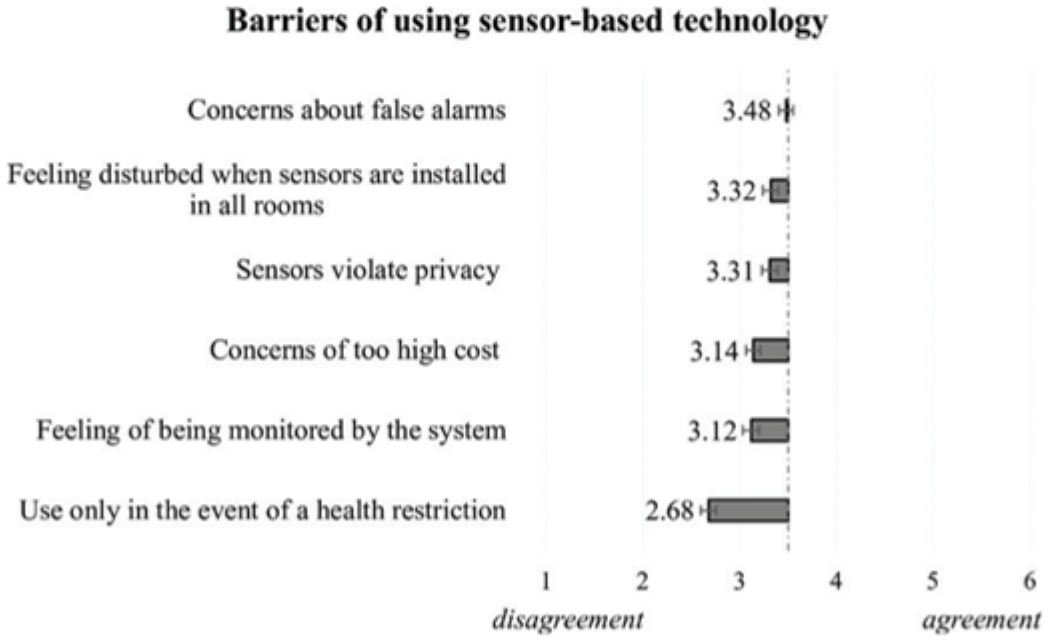


Fig 7.5 Perceptions of barriers when using sensor-based technology.
https://doi.org/10.1371/journal.pone.0269642.g005

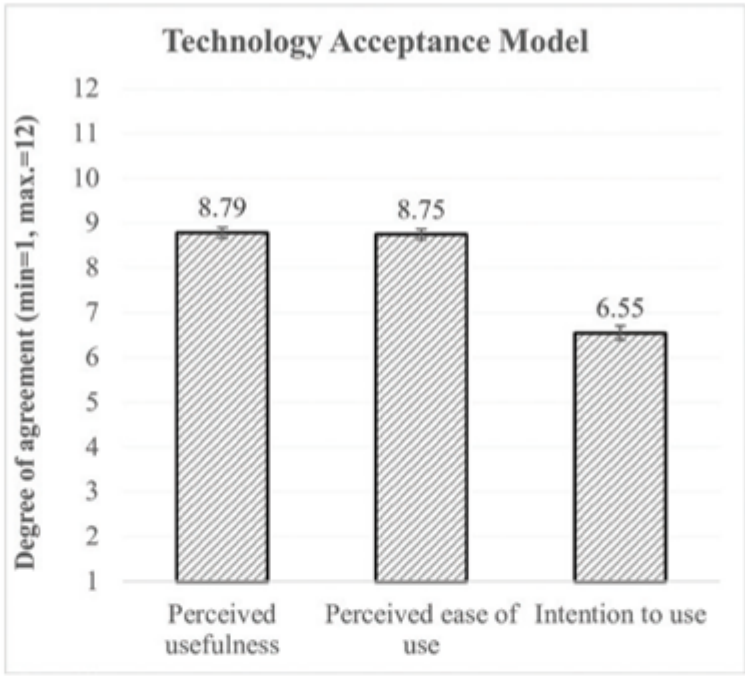


Fig 7.6 Evaluations of the technology acceptance criteria according to TAM [18].
https://doi.org/10.1371/journal.pone.0269642.g006

Perceptions of privacy

In the next step of the statistical analysis, we consider the aspect of privacy and its role for the use of ambient sensor-based applications. For this purpose, we firstly asked the survey participants about the importance of personal privacy in using sensor-based lifelogging in general as well as in terms of individual activities in different rooms at home. Secondly, we analyzed which role plays health condition in the perception of privacy. And finally, we reveal in which relation perceptions of privacy are connected with the technology acceptance of the assistive technology.

Privacy: A curse or a blessing?. In our questionnaire, participants were asked to assess their privacy related to the use of sensor-based technology in their private homes. The majority of the respondents agreed that privacy protection should have the first priority ($M = 4.39$, $SD = 1.4$) and they also agreed that they would use sensor-based lifelogging only as long as their privacy is protected ($M = 4.38$, $SD = 1.4$). As it is shown in Fig 7.7, participants were overall relatively neutral towards recording for their own purposes ($M = 3.5$, $SD = 1.5$).

But which activities, or which areas in the private environments, cause the reluctance by users regarding the addressed technology? According to the results presented in Fig 7.8, the use of sensor-based technology in the home environment was truly critical only in two rooms, which insinuate a high intimacy requirement anyway. These rooms are the bathroom ($M = 76.4$; $SD = 27.9$, where 100 is the most critical value) and the bedroom ($M = 76.3$; $SD = 26.9$) showing almost identical means resulting from the assessments of the possible recordings of personal activities in these rooms. As less critical respondents perceived the stay in the kitchen ($M = 38.1$; $SD = 29.3$) as well as duration ($M = 44.4$; $SD = 30.7$) and frequency of stays ($M = 43.3$; $SD = 33.1$), and the timing of entering and leaving different rooms ($M = 45.4$; $SD = 33.1$). The other examined activities were perceived as rather neutral.

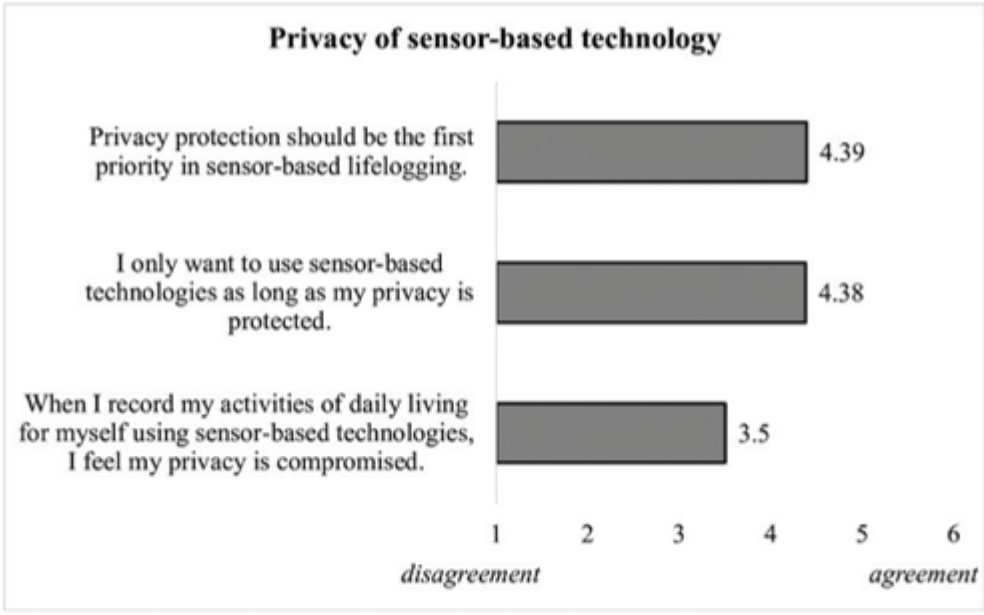


Fig 7.7 Evaluations of privacy in the context of sensor-based technology.

https://doi.org/10.1371/journal.pone.0269642.g007

Fig 7.8 Wariness about privacy issues when using sensor-based systems.

https://doi.org/10.1371/journal.pone.0269642.g008

What role does privacy still play in the use of the technology when it comes to health?. Or maybe we should just ask the other way around: Which role does health condition play in the perception of privacy? Individual health was found to play a decisive role in the acceptance and use of health information technologies (e.g., [41, 42]). For this reason, the current study additionally examined how participants assess the relevance, and thus possibly the disclosure, of their health in relation to the privacy issues resulting from the logging of sensor-based data. To avoid biased evaluations we randomly used statements, which are both privacy-affirming and privacy-denying in the presence of a medical condition or physical limitation; for details see Fig 7.9.

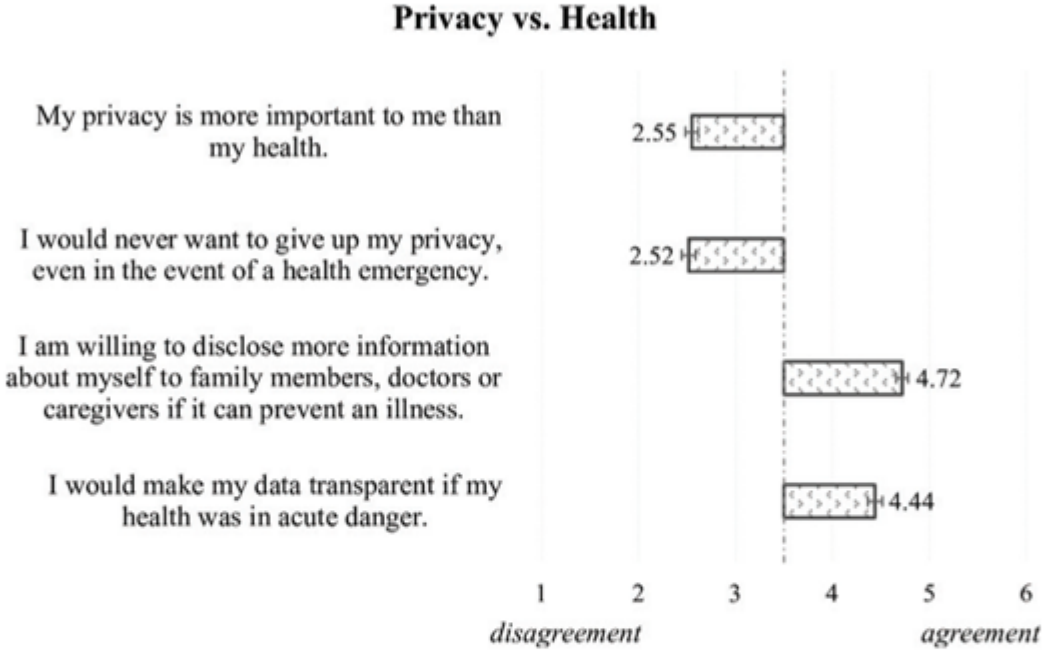


Fig 7.9 The relevance of health in the perception of privacy when using sensor-based technology.
https://doi.org/10.1371/journal.pone.0269642.g009

In the comparison of how important privacy is in relation to one's own health, the outcomes reveal that health clearly takes the first position. Survey respondents decisively rejected the higher importance of privacy ($M = 2.5$, $SD = 1.2$) and their persistence as to privacy protection in the light of a health emergency ($M = 2.5$, $SD = 1.2$). In contrast, according to the results, they would disclose the necessary information to important others when this would prevent an illness ($M = 4.7$, $SD = 1.1$), and they would make their data transparent in the situation of acute danger ($M = 4.4$, $SD = 1.3$).

Relations between technology acceptance and the perceptions of privacy

For the final step of the statistical analyses, the question remains to what extent perceptions of privacy determine the acceptance of the use of sensor-based technology in home environments. For this purpose, we used the previously described construct 'privacy and the health data' (see Methods section) in connection with the key criteria of technology acceptance model.

A stepwise regression analysis revealed that privacy is significantly related to the accepted use of the technology. The model was statistically significant, $F(1,299) = 104.52$, $p \leq .001$, and accounted for approximately 26 percent of the variance of acceptance ($R^2 = .260$, adjusted $R^2 = .257$). The privacy as a

predictor received the weight of β = .51 (t(299) = 10.22, p≤.001), making a strong unique contribution to explain the technology acceptance of ambient sensors.

As can be seen in Table 7.2, privacy explained the variance of the perceived usefulness for the most (25.4%), followed by the intention to use (15.6%) and the perceived ease of use (7.8%), thus making its role as a significant predictor for the acceptance evident.

Table 7.2 Results of a regression analysis for the acceptance of sensor-based technology in home environments (*N* = 312; ***p*≤.001; VIF = variance inflation factor < 10).

	Predictor	Adj. R^2	β	t	VIF	ANOVA
Acceptance of sensor-based technology	Privacy and/vs. health	25.7%	.51	10.22**	1.0	F(1,299) = 104.52, p≤.001
Perceived usefulness		25.4%	.51	10.11**	1.0	F(1,297) = 102.22, p≤.001
Intention to use		15.6%	.40	7.40**	1.0	F(1,291) = 54.73, p≤.001
Perceived ease of use		7.8%	.28	5.10**	1.0	F(1,295) = 25.97, p≤.001

https://doi.org/10.1371/journal.pone.0269642.t002

Discussion

Smart technologies are becoming an increasingly important part of our lives. Miniaturized computer technology makes objects, in which it finds its way, part of a connected information and communication system and has the potential to continuously log personal data [43]. Sensor-based technologies that are not only available in the form of wearable devices but can also be implemented in the environment and be networked via mobile communications, enable to collect data from the surroundings without the user's active intervention and can exchange data, discreetly supporting the users in many areas of life.

The potentially life-enriching benefits of sensor-based lifelogging technologies, as described in the presented study, can be useful for the aging society and can improve autonomy as well as medical treatment and care of their users. Nevertheless, feasibility alone does not guarantee success in the implementation and sustainable use of such assistive lifelogging applications. To fully exploit the potential offered by these technologies and to ensure a long-term implementation, it is thus necessary to investigate possible acceptance issues from the perspective of (potential) users. To gain insights into the significant influences on the adoption of sensor-based lifelogging, we empirically investigated concrete acceptance factors as well as privacy issues—previously identified as critical in the lifelogging context—and the interrelationship between these variables. In the following, we discuss the findings and outline the need for further research in this context, considering at the same time the limitations of the presented study.

User-centered preferences and acceptance-relevant considerations

According to our results, German adults eminently appreciate the benefits of sensor-based lifelogging: In the ranking, the highest approvement is assigned to the benefit of emergency functions, but the sense of safety and the ability of an independent living in the own four walls, that are ensured by the use of the technology, are also highly appreciated. These outcomes resonate with previous research in assistive lifelogging (e.g., [26, 29, 31]). On the side of the perceived barriers when using sensor-based applications, the participants mostly disagreed with the selected items. The highest uncertainty is noticeable with respect to the possible

false alarms; invasion of privacy as well as the saturation of sensors in all living spaces could also pose potential drawbacks for the acceptance. However, from the perception of benefits and barriers, opinions about the assistive lifelogging technologies in private home environments are basically positive.

To theoretically substantiate the aspect of technology acceptance, in our study we additionally examined the key acceptance criteria according to the renowned Technology Acceptance Model [18]. The survey results revealed that the potential users perceive the sensor-based lifelogging as useful and easy to use, nevertheless they intend to use it only hesitantly. This reluctance of using the ambient sensors may be due to the fact that these applications are not yet well-known and that their functionalities (e.g., drawing conclusions about daily habits and activities, identifying events that deviate from the daily pattern) are not immediately obvious to the respondents. On the other hand, assuming that the respondents were familiar with the benefits provided by the sensor-based lifelogging, as the sample was on average at just under 33 years still quite young, they may not have realized the added value for themselves—at least not at the current time in their lives. Thus, answering to the previously defined research questions (RQ1, RQ2) on the basis of our findings, we can state that the general acceptance of sensor-based technology is high. However, although the potential users seem to be guided more by the perceptions of benefits than drawbacks brought by the technology, and they clearly perceive its usefulness, the willingness to an immediate utilization is rather reserved.

In the next step, we focused on results concerning personal privacy, since in previous research the loss of privacy was identified as an important barrier towards the adoption of technologies in the AAL context (e.g., [39, 40]). Our findings revealed a high need for privacy protection: The majority of our sample assigned the first priority to the protection of their privacy when using sensor-based lifelogging and they agreed on using that technology only if their privacy is properly protected. When asked about specific rooms and types of sensor-logged activities, however, it turned out that only recordings in the bathroom and bedroom are perceived as critical, while other rooms are judged to be quite neutral for the lifelog. Nonetheless, the attitude about the highest priority for personal privacy changes when health issues come into play. Our survey respondents decisively rejected the higher importance of privacy in case of a health emergency: They would rather provide information needed to prevent an illness and disclose their health data in the situation of acute danger. These findings show that the perceptions of privacy are clearly situation- and context-dependent, playing an important role in the adoption of sensor-based ambient technologies (RQ3).

Furthermore, as for the acceptance of such technology in home environments we found, still considering the aspect of health, that privacy protection accounts to 26% for the willingness to use, and thus for an accepted adoption of, assistive sensor-based lifelogging technologies. This result of our study indicates that besides other factors, presumably having a decisive impact on acceptance, privacy protection significantly influences the adoption of such assistive technologies (RQ4).

With this in mind, we can state that potential users generally show an accepting attitude towards assistive sensor-based logging in their home environments and require high privacy protection standards for their personal data. Nevertheless, there is also the willingness to lower these requirements in the event of health-related changes in life situations.

Limitations and future work

In addition to the relevant findings that this study has revealed in the context of a specific technology and use, some limitations and, in particular, the need for future research should be also addressed.

From the methodological point of view, an important limitation of the results pertains to the relatively young sample. Since young users predominantly have been shown to have greater technical expertise [44, 45] and perceived familiarity with new technology [46], this is probably also true, at least to some extent, for the

investigated sensor-based lifelogging. In future studies, therefore, the opinions regarding acceptance and privacy requirements are to be explored primarily from the perspective of older users, as they are the main target user group for this kind of assistive technology. Here, it is even conceivable that there are significant differences between healthy aging and chronically ill elders, which should be deliberately taken into account in the studies.

There are also other arguments in favor of focusing on older users in the described research context: Resulting from our findings, the willingness to an immediate use of sensor-based technology is rather reserved. At this point, we can only hypothesize the reasons for such an attitude, especially because of the resulting high general acceptance. Future research has thus to find out which factors or circumstances caused this finding. Furthermore, the question emerges whether this is a generally valid outcome or it is due to the specific characteristics of the participating respondents. It is conceivable that information derived from the everyday life of older persons or those in need of care and the conclusions drawn from the technology about their habits and activities will be perceived as very beneficial in the group of older users. Thus, future work should evaluate opinions of predominantly older users, who are confronted with a necessity of assistance in some areas of their life.

Conclusions

The results of our empirical study demonstrate that the participants as potential users generally represent an accepting mindset towards ambient sensors that log different parameters in their living environments. The advantages and the perceived usefulness of the technology strongly outweigh the possible weaknesses. However, we must also note that the intention to use sensor-based lifelogging technologies is still relatively low. Technology acceptance is furthermore determined—to a non-trivial extent—by privacy protection, and users require a high privacy protection by default; even though they are willing to accept more transparency of personal data in the event that they suffer a health decline.

Supporting Information

S1 Data.
(CSV)

Acknowledgments

We would like to thank all participants for sharing their opinions on using sensor-based technologies in home environments.

Author Contributions

Conceptualization: Wiktoria Wilkowska, Julia Offermann.
Data curation: Wiktoria Wilkowska, Julia Offermann.
Formal analysis: Wiktoria Wilkowska.
Funding acquisition: Susanna Spinsante, Martina Ziefle.
Investigation: Wiktoria Wilkowska, Julia Offermann.
Methodology: Wiktoria Wilkowska, Julia Offermann.
Project administration: Susanna Spinsante, Martina Ziefle.
Resources: Wiktoria Wilkowska, Julia Offermann.
Software: Wiktoria Wilkowska, Julia Offermann.
Supervision: Susanna Spinsante.

Validation: Wiktoria Wilkowska, Julia Offermann.
Visualization: Wiktoria Wilkowska.
Writing – original draft: Wiktoria Wilkowska, Julia Offermann, Susanna Spinsante, Angelica Poli.
Writing – review & editing: Wiktoria Wilkowska, Julia Offermann, Susanna Spinsante, Angelica Poli.

References

1. Selke S. Lifelogging: Digital self-tracking and Lifelogging-between disruptive technology and cultural transformation. Springer; 2016. https://doi.org/10.1007/978-3-658-13137-1
2. Rashidi P, Mihailidis A. A Survey on Ambient-Assisted Living Tools for Older Adults. IEEE J Biomed Health Inform. 2013; 17(3): 579–590. https://doi.org/10.1109/jbhi.2012.2234129 PMID:24592460
3. Blackman S, Matlo C, Bobrovitskiy C, Waldoch A, Fang ML, Jackson P, et al. Ambient assisted living technologies for aging well: A scoping review. Int J Intell Syst. 2014; 25(1): 55–69. https://doi.org/10.1515/jisys-2014-0136
4. Poli A, Cosoli G, Scalise L, Spinsante S. Impact of wearable measurement properties and data quality on ADLs classification accuracy. IEEE Sens. J. 2020; 21(13): 14221–14231. https://doi.org/10.1109/JSEN.2020.3009368
5. Rashidi P, Cook DJ. Keeping the resident in the loop: Adapting the smart home to the user. IEEE Trans Sys Man Cybern A: Sys Hum. 2009; 39(5): 949–959. https://doi.org/10.1109/TSMCA.2009.2025137
6. Mubashir M, Shao L, Seed L. A survey on fall detection: Principles and approaches. Neurocomputing. 2013; 100: 144–152. https://doi.org/10.1016/j.neucom.2011.09.037
7. Spinsante S, Poli A, Mongay Batalla J, Krawiec P, Dobre C, Bǎjenaru L, et al. Clinically-validated technologies for assisted living. J Ambient Intell Humaniz Comput. 2021 Aug: 1–22. https://doi.org/10.1007/s12652-021-03419-y
8. Hayes TL, Abendroth F, Adami A, Pavel M, Zitzelberger TA, Kaye JA. Unobtrusive assessment of activity patterns associated with mild cognitive impairment. Alzheimers Dement. 2008; 4(6): 395–405. https://doi.org/10.1016/j.jalz.2008.07.004 PMID:19012864
9. Suzuki T, Murase S, Tanaka T, Okazawa T. New approach for the early detection of dementia by recording in-house activities. Telemed E-Health. 2006; 13(1): 41–44. https://doi.org/10.1089/tmj.2006.0033 PMID:17309353
10. Vimalkumar M, Sharma SK, Singh JB, Dwivedi YK. 'Okay google, what about my privacy?': User's privacy perceptions and acceptance of voice based digital assistants. Comput Hum Behav. 2021; 120: 106763. https://doi.org/10.1016/j.chb.2021.106763
11. Wilkowska W, Offermann-van Heek J, Colonna L, Ziefle M. Two faces of privacy: Legal and human-centered perspectives of lifelogging applications in home environments. In: Gao Q, Zhou J, editors. Human aspects of IT for the aged population. Healthy and active aging. Lecture notes in computer science (vol. 12208). Springer; 2020: pp. 545–564. https://doi.org/10.1007/978-3-030–50249–2_39
12. Mihaildis A, Colonna L. A methodological approach to privacy by design within the context of lifelogging technologies. Rutgers Computer & Tech. LJ; 2020; 46: 1.
13. Gurrin C, Albatal R, Joho H, Ishii K. A Privacy by Design Approach to Lifelogging. In: O'Hara K, Nguyen M-HC, Haynes P, editors. Digital Enlightenment Yearbook 2014: Social Networks and

Social Machines, Surveillance and Empowerment. IOS Press.; 2014: pp. 49–73. https://doi.org/10.3233/978-1-61499-450-3-49

14. Biniok P, Hülsmann I. 21st Century Men and the Digital Amalgamation of Life. In: Selke S, editor. Lifelogging: Digital self-tracking and Lifelogging—Between disruptive technology and cultural transformation. Springer Fachmedien Wiesbaden; 2016. (pp. 81–108). https://doi.org/10.1007/978-3-658-13137-1_5

15. Climent-Pérez P, Spinsante S, Mihailidis A, Florez-Revuelta F. A review on video-based active and assisted living technologies for automated lifelogging. Expert Syst Appl. 2020; 139: 1–19. https://doi.org/10.1016/j.eswa.2019.112847

16. Mainetti L, Patrono L, Secco A, Sergi I. An IoT-aware AAL System to Capture Behavioral Changes of Elderly People. J Commun Softw Syst. 2017; 13(2): 68–76. https://doi.org/10.24138/jcomss.v13i2.374

17. Gurrin C, Smeaton AF, Doherty AR. LifeLogging: Personal Big Data. Found Trends Inf Ret. 2014; 8(1): 1–125. https://doi.org/10.1561/1500000033

18. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly. 1989; 13: 319–340. https://doi.org/10.2307/249008

19. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: Toward a unified view. MIS Quarterly. 2003; 425–478. https://doi.org/10.2307/30036540

20. Venkatesh V, Thong JY, Xu X. Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Quarterly. 2012; 36(1): 157–178. https://doi.org/10.2307/41410412

21. Turner M, Kitchenham B, Brereton P, Charters S, Budgen D. Does the technology acceptance model predict actual use? A systematic literature review. Inf Softw Technol. 2010; 52(5): 463–479. https://doi.org/10.1016/j.infsof.2009.11.005

22. Venkatesh V, Davis FD. A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. Manag Sci. 2000; 46(2): 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

23. Peek ST, Wouters EJ, Van Hoof J, Luijk KG, Boeije HR, Vrijhoef HJ. Factors influencing acceptance of technology for aging in place: a systematic review. Int J Med Inform. 2014; 83(4): 235–248. https://doi.org/10.1016/j.ijmedinf.2014.01.004 PMID:24529817

24. Haesner M, O'Sullivan JL, Gövercin M, Steinhagen-Thiessen E. Requirements of older adults for a daily use of an internet-based cognitive training platform. Inform Health Soc Care. 2015; 40(2): 139–153. https://doi.org/10.3109/17538157.2013.879149 PMID:24725153

25. Sharon T. Self-tracking for health and the quantified self: Re-articulating autonomy, solidarity, and authenticity in an age of personalized healthcare. Philos Technol. 2017; 30(1): 93–121. https://doi.org/10.1007/s13347-016-0215-5

26. Van Hoof J, Kort HS, Rutten PG, Duijnstee MSH. Ageing-in-place with the use of ambient intelligence technology: Perspectives of older users. Int J Med Inform. 2011; 80(5): 310–331. https://doi.org/10.1016/j.ijmedinf.2011.02.010 PMID:21439898

27. Biermann H, Himmel S, Offermann-van Heek J, Ziefle M. User-specific concepts of aging–A qualitative approach on AAL- acceptance regarding ultrasonic whistles. In: Zhou J, Salvendy G,

editors. Human aspects of IT for the aged population. Applications in health, assistance, and entertainment. Lecture notes in computer science (vol. 10927). Springer; 2018. pp. 231–249. https://doi.org/10.1007/978–3-319-92037-5_18

28. Lorenzen-Huber L, Boutain M, Camp LJ, Shankar K, Connelly KH. Privacy, technology, and aging: A proposed framework. Ageing Int. 2011; 36(2): 232–252. https://doi.org/10.1007/s12126-010-9083-y

29. Jaschinski C, Allouch SB. Listening to the ones who care: exploring the perceptions of informal caregivers towards ambient assisted living applications. J Ambient Intell Humaniz Comput. 2019; 10(2): 761–778. https://doi.org/10.1007/s12652-018-0856-6

30. Sun H, De Florio V, Gui N, Blondia C. The missing ones: Key ingredients towards effective ambient assisted living systems. J Ambient Intell Smart Environ. 2010; 2(2): 109–120. https://doi.org/10.3233/AIS-2010-0062

31. Steele R, Lo A, Secombe C, Wong YK. Elderly persons' perception and acceptance of using wireless sensor networks to assist healthcare. Int J Med Inform. 2009; 78(12): 788–801. https://doi.org/10.1016/j.ijmedinf.2009.08.001 PMID:19717335

32. Lidynia C, Schomakers EM, Ziefle M. What are you waiting for?–Perceived barriers to the adoption of fitness-applications and wearables. In: Ahram T, editor. Advances in intelligent systems and computing: Vol. 795. Advances in human factors in wearable technologies and game design. Springer; 2018: pp. 41–52. https://doi.org/10.1007/978-3-319-94619-1_5

33. Himmel S, Ziefle M. Smart home medical technologies: users' requirements for conditional acceptance. i-com. 2016; 15(1): 39–50. https://doi.org/10.1515/icom-2016-0007

34. Offermann-van Heek J, Schomakers EM, Ziefle M. Bare necessities? How the need for care modulates the acceptance of ambient assisted living technologies. Int J Med Inform. 2019; 127: 147–156. https://doi.org/10.1016/j.ijmedinf.2019.04.025 PMID:31128827

35. Gelonch O, Ribera M, Codern-Bové N, Ramos S, Quintana M, Chico G, et al. Acceptability of a lifelogging wearable camera in older adults with mild cognitive impairment: a mixed-method study. BMC Geriatr. 2019; 19: 110. https://doi.org/10.1186/s12877-019-1132-0 PMID:30991948

36. Goddard M. The EU General Data Protection Regulation (GDPR): European regulation that has a global impact. Int J Mark Res. 2017; 59(6): 703–705. https://doi.org/10.2501/IJMR-2017-050

37. Padilla-López J, Chaaraoui A, Gu F, Flórez-Revuelta F. Visual privacy by context: Proposal and evaluation of a level-based visualisation scheme. Sensors. 2015; 15(6): 12959–12982. https://doi.org/10.3390/s150612959 PMID:26053746

38. Wilkowska W, Offermann-van Heek J, Florez-Revuelta F, Ziefle M. Video Cameras for Lifelogging at Home: Preferred Visualization Modes, Acceptance, and Privacy Perceptions among German and Turkish Participants. Int J Hum–Comput Interact. 2021; 37 (15): 1436–1454. https://doi.org/10.1080/10447318.2021.1888487

39. Schomakers E-M, Ziefle M. Privacy Perceptions in Ambient Assisted Living. In: Proceedings of the 5th International Conference on Information and Communication Technologies for Ageing Well and E-Health (ICT4AWE 2019). Scitepress: pp. 205–212. https://doi.org/10.5220/0007719802050212

40. Jaschinski C, Allouch SB, Peters O, van Dijk J. The Influence of Privacy on the Acceptance of Technologies for Assisted Living. In: Gao Q, Zhou J, editors. Human Aspects of IT for the Aged

Population. Healthy and Active Aging. HCII 2020. Lecture Notes in Computer Science (vol 12208). Springer, Cham; 2010: pp. 463–473 https://doi.org/10.1007/978-3-030-50249-2_33

41. Shen N, Bernier T, Sequeira L, Strauss J, Silver MP, Carter-Langford A, et al. Understanding the patient privacy perspective on health information exchange: a systematic review. Int J Med Inform. 2019; 125: 1–12. https://doi.org/10.1016/j.ijmedinf.2019.01.014 PMID:30914173

42. Park YJ, Shin DD. Contextualizing privacy on health-related use of information technology, Comput Hum Behav. 2020; 105: 106204. https://doi.org/10.1016/j.chb.2019.106204

43. Hilty LM, Behrendt S, Binswanger M, Bruinink A, Erdmann L, Fröhlich J, et al. Das Vorsorgeprinzip in der Informationsgesellschaft: Auswirkungen des Pervasive Computing auf Gesundheit und Umwelt [The Precautionary Principle in the Information Society: Effects of Pervasive Computing on Health and the Environment]. TA-Swiss, Zentrum für Technologiefolgen-Abschätzung, Bern; 2003 Jan. ISBN: 3-908174-06-6

44. Czaja SJ, Charness N, Fisk AD, Hertzog C, Nair SN, Rogers WA, et al. Factors predicting the use of technology: findings from the Center for Research and Education on Aging and Technology Enhancement (create). Psychol Aging. 2006; 21(2): 333–352. https://doi.org/10.1037/0882-7974.21.2.333 PMID:16768579

45. Lee C, Coughlin JF. PERSPECTIVE: Older adults' adoption of technology: an integrated approach to identifying determinants and barriers. J Prod Innov Manag. 2015; 32(5): 747–759. https://doi.org/10.1111/jpim.12176

46. Brown SA, Venkatesh V. Model of adoption of technology in households: A baseline model test and extension incorporating household life cycle. MIS Quarterly. 2005; 29(3): 399–426. https://doi.org/10.2307/25148690

7.7 Case Study 2

RESEARCH ARTICLE

Blockchain technology in healthcare: A systematic review

Huma Saeed[1], Hassaan Malik [1,2]*, Umair Bashir[1], Aiesha Ahmad[1], Shafia Riaz[1], Maheen Ilyas[1], Wajahat Anwaar Bukhari[1], Muhammad Imran Ali Khan[1]

1 Department of Computer Science, National College of Business Administration & Economics Lahore, Multan, Pakistan, 2 Department of Computer Science, University of Management and Technology, Lahore, Pakistan

* *f2019288004@umt.edu.pk*

Abstract

Blockchain technology (BCT) has emerged in the last decade and added a lot of interest in the healthcare sector. The purpose of this systematic literature review (SLR) is to explore the potential paradigm shift in healthcare utilizing BCT. The study is compiled by reviewing research articles published in nine well-reputed venues such as IEEE Xplore, ACM Digital Library, Springs Link, Scopus, Taylor & Francis, Science Direct, PsycINFO, Ovid Medline, and MDPI between January 2016 to August 2021. A total of 1,192 research studies were identified out of which 51 articles were selected based on inclusion criteria for this SLR that presents the modern information on the recent implications and gaps in the use of BCT for enhancing the healthcare procedures. According to the outcomes, BCT is being applied to design the novel and advanced interventions to enrich the current protocol of managing, distributing, and processing clinical records and personal medical information. BCT is enduring the conceptual development in the healthcare domain, where it has summed up the substantial elements through better and enhanced efficiency, technological innovation, access control, data privacy, and security. A framework is developed to address the probable field where future researchers can add considerable value, such as data protection, system architecture, and regulatory compliance. Finally, this SLR concludes that the upcoming research can support the pervasive implementation of BCT to address the critical dilemmas related to health diagnostics, enhancing the patient healthcare process in remote monitoring or emergencies, data integrity, and avoiding fraud.

Introduction

Healthcare is a system that includes 3 main components: (i) Main suppliers of services for medical treatment, For instance, doctors, nurses, technicians, and hospital administrations (ii) Emergency related services [1–4], and (iii) Health and health-oriented service users, specific patients. In the current study, to encourage, preserve or restore the health of beneficiaries, we examine the health maintenance to include technology-based remotely controlling services increased by constituent service providers [5–9]. In the medical field, every year, there are more security and privacy breaches, in 2017, more than 300 breaches were reported, and up to 37 million records were affected during 2010–2017 [10, 11]. The growing digitization of medical care has advanced the acknowledgment of issues about secure storage, accessing of patients' medical records, ownership, and medical data from associated sources [12–16]. Blockchain is recommended as a method of addressing critical issues faced by healthcare, for instance, protected sharing of health records and adherence to data privacy laws [17–19].

https://doi.org/10.1371/journal.pone.0266462 Used with CC 4.0 license

Blockchain is a particular type of database that can be managed by the network of authenticated members or nodes [13] and stores immutable information blocks that can be strongly exchanged without interference by third parties [10]. With cryptographic signatures and the use of consensus algorithms which are implemented as key enablers in their application, data is stored and registered [20]. The capability of preserving data is a major aim for using the BCT particularly in healthcare [21], which is subject to massive sharing and dissemination of a significant amount of data [7]. In different stages, the development of blockchain technology, as well as its application in various contexts, had been materialized. The first phase of blockchain development was focused on cryptocurrencies, while the second focused on the use of smart contracts in industries like real estate and finance [11, 22]. The 3rd generation of evolution concentrated on employing blockchain in non-financial areas including government, culture [23], and healthcare space [22, 24]. Also, powered by revolutionary technical features such as data immutability [25], with the introduction of artificial intelligence, blockchain technology is having its 4th generation of evolution [26]. This asserted diversity in Blockchain's application spectrum can be attributed to its ability to build decentralized [27] and trustless transaction environments [28]. As blockchain can tackle serious issues, such as automated claim authentication [9] and public health management [29], the healthcare sector is a prime choice for the application of blockchain technology [30–32]. This technology allows patients to keep personal data and determine with whom this can be shared, thus resolving current data ownership, and sharing issues [28, 33]. At the same time, it allows recorded data to be integrated, modified, shared safely, and retrieved on time by relevant authorities using consensus protocols [31]. This is a significant benefit of the use of this technology in the healthcare system, as existing procedures need third parties to store the data [10]. Finally, because of possible human error, blockchain could potentially add accountability to data management processes [34] further decreasing the risks of mishandling or misusing recorded data [31]. Given the optimistic connotations of the effects of blockchain on social and business change, in contrast to previously defined expectations, it appears to be a discussion regarding its basic and derived advantages. A recent study indicates that while organizations will make substantial investments in the future in adopting blockchain-based technology due to a widespread perception that the advantages could be over-hype, they will probably accept a cautiously pragmatic approach [35]. It can be said that this technology has yet to fulfill its expectations [36], a fact that can be due to the prevalent adoption of block chain, particularly about regulatory barriers, to certain challenges [31]. The general public and specific users, for instance, patients or physicians are not acquainted with the way blockchain works, the technological features, or its advantages for data processing is another significant obstacle in promulgating the implementation of blockchain [35]. Suggest that it may take a considerable time for this technology to establish all anticipated stages of business transformation mainly because of the organizational, social, and implementation challenges, for example, security issues or governance reasons [22, 31]. This could also be exacerbated by general confusion regarding the use of blockchain regarding legal enforcement and regulations of the government. Current research focuses on supporting blockchain operational growth and speed-up its prevalence by overcoming these barriers.

However, previous studies have made little attempt to comprehensively summarize the existing knowledge by using SLRs [9–13]. For example, bibliometric techniques were used by [10] to provide a summary of blockchain research patterns and components related to the implementation of blockchain in the field of healthcare. In [9] the different blockchain platforms have been developed to deploy blockchain in healthcare. The study [11] addressed different examples of the implementation in the healthcare of blockchain technology, the problems, and their potential solutions. In diverse contexts where this technology was implemented [12], addressed design choices and tradeoffs made by the researchers. The research studies of [13, 14] have discussed the Blockchain-based applications throughout numerous industries and addressed many contexts of use for this technology in a broad manner. Recently [14] reviewed 39 studies to present an overview on common channels and other areas where blockchain technology is utilized for healthcare enhancement. Although these systematic literature reviews have a contribution to the extent of knowledge, their emphasis

https://doi.org/10.1371/journal.pone.0266462

has been mainly on synthesizing or delineating blockchain technology patterns and areas [10, 11, 13, 14, 16]. However, researchers will get benefit from a concentrated discussion on the implications of its adoption [15], along with concrete obstacles and areas for progress for advancing the field, due to the reach and diversity of previous blockchain studies [11]. Through assimilating existing information and describing focus areas that require considerable academic attention, review-based research will assist in meeting these needs [11, 16–19]. As a result of this necessity, we perform an SLR on the blockchain technology application. This SLR presents a valuable overview of ongoing research, gaps in current knowledge, and future avenues of research as well. The contribution of this study is in two ways, this research adds to the emerging blockchain literature in healthcare. First regarding their implementation areas, restrictions, and recommendations, it offers an advanced and thematically ordered classification of previous literature. Second, we propose a synthesizing process according to the results of the SLR to detail possible topics that need academic attention to further update the existing body of literature.

The present study is organized as follows: In Section 2, we provide a thorough description of the research method utilized to search, screen, and select the literature. In Section 3, we present relevant review works that have been conducted in the field of health care using blockchain technology and discuss all the papers that have been selected, focusing on their main findings, and highlighting research gaps for future research. Finally, in Section 4we conclude this study.

Methodology

SLRs always provide a thorough understanding of literature as it presents a complete and systematized review meeting all standard protocols in it [18, 37–39]. SLRs also help in the understanding of current information gaps and, as a result, the discovery of potential research avenues [19].

Research questions

We conduct this SLR by addressing the following research questions (RQs).

RQ1: What is the advanced profile used for the employment of blockchain in the healthcare domain? The purpose of this research question is to identify the number of research papers issued every year, the average citation received on research papers yearly, and academic contribution on the subject by Journals, publishing houses, and community.

RQ2: What are the major healthcare domains where blockchain technology has been implemented? The purpose of this question is to identify the contexts in which blockchain technology has shown significant outcomes in healthcare.

RQ3: What are the existing problems and constraints raised by the previous studies in the healthcare field using blockchain technology? The motivation behind this question is to identify existing problems and issues of blockchain technology in the healthcare field based on results, limitations, and conclusions of previous research studies.

RQ4: What are the potential healthcare avenues that would benefit from blockchain technology implementation? The purpose of this question is to identify growing gaps and prospects of the future research agenda

Research objectives

The research objectives (ROs) of the article herein presented are the following:

RO1: Establishing an archive of work that relates a wide topic about Blockchain in healthcare and offers an open dataset about Blockchain for all other researchers.

https://doi.org/10.1371/journal.pone.0266462

RO2: Identify a more focused set of studies that have used blockchain technology in healthcare applications.

RO3: Identify problems and constraints discussed in the healthcare field using blockchain technology.

RO4: Characterize existing solutions in the field of blockchain in healthcare and clarify the similarities and differences between them using a characterization framework.

Research strategy

Nine databases—IEEE Xplore, ACM Digital Library, Springs Link, Scopus, Taylor & Francis, Science Direct, PsycINFO, Ovid Medline, MDPI—are recognized by previous studies as standard data sources of research papers about health informatics [40]. Reviewed papers have been outlined for understanding the research status of applying blockchain in health care. For the right database search the three keyword combinations existed as—"blockchain and Healthcare", or "medical Health" or Medical Management or Health Management. The above keywords were extracted from an article of previous literature i.e. SLRs) using similar keywords such as blockchain and medical healthcare.

Study selection

The selection process aimed to find the articles that are the most relevant to the objective of this SLR. If there was the same paper in more than one source, as per our research, it was considered only once. The content of the papers chosen for the final sample was evaluated [39, 41] to make sure that the findings of the present SLR produced clear results and that is not biased. For reaching a consensus of final inclusion or exclusion, two of the researchers finalized the evaluation. After completing this, the discrepancies of individual assessments were addressed through discussion. A third author was engaged in analysis and debate in situations where the two writers did not find consensus. After the papers were found, the first move was to delete redundant titles and those which are not connected in scrutiny. The standards for inclusion were limited to the hunt for String, and a study conducted by at least one of the following criteria for exclusion (EC) is omitted:

Inclusion criteria (ICs).

IC1: Studies are released any time on or before August 2021.

IC2: Studies are limited to the journal, conference, report, workshop, and symposium articles only.

IC3: Availability of complete texts in digital databases.

IC4: Proposed models or frameworks present.

Exclusion criteria (EC).

EC1: Exclude duplicated studies.

EC2: Eliminate preview, book chapters, magazines, thesis, monographs, and interview-based articles.

EC3: Exclude studies based on quality evaluation criteria.

EC4: Studies written in a language other than English.

The choice of papers was based on clear above discussed criteria for inclusion and exclusion. Below Fig 7.10 has been developed through the aspiration from the PRISMA diagram [42]. Fig 7.10 shows the study selection process.

https://doi.org/10.1371/journal.pone.0266462

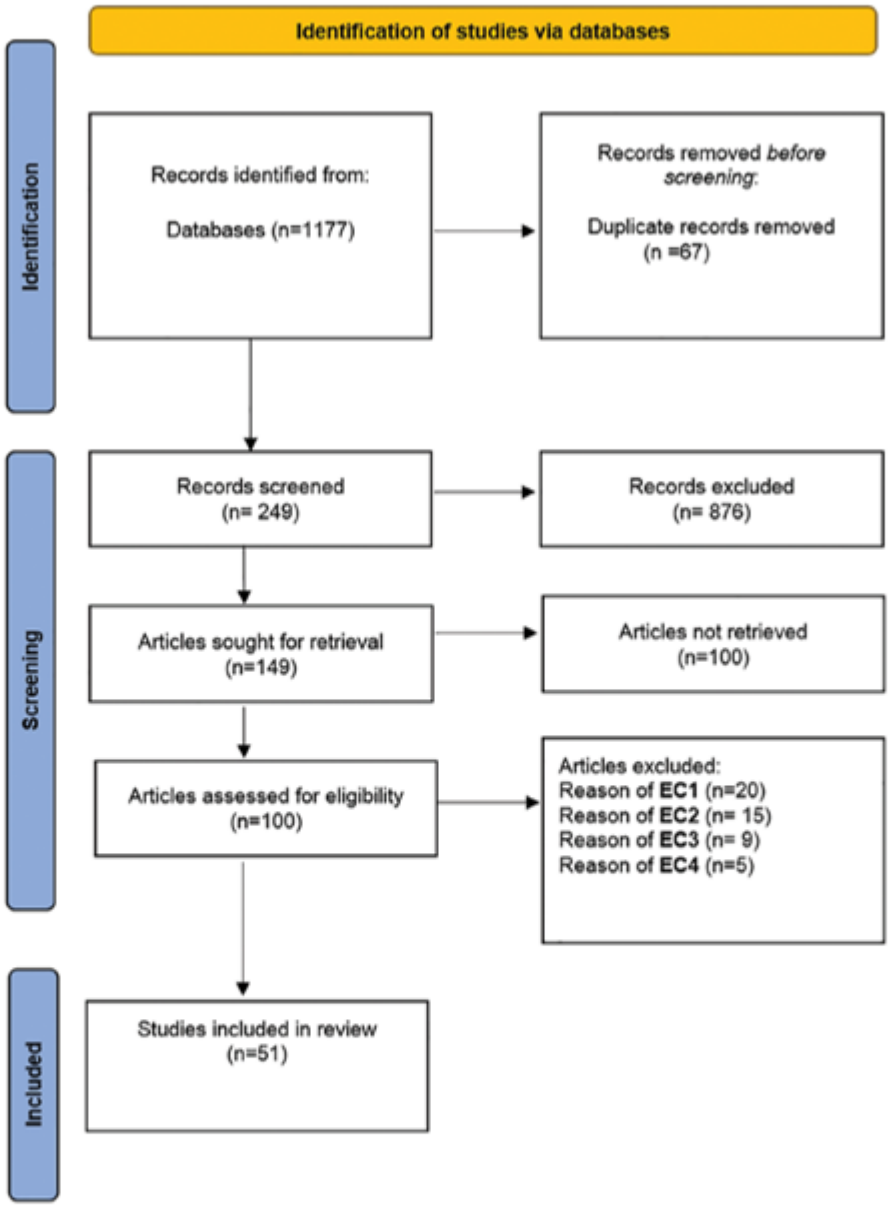


Fig 7.10 **PRISMA flow chart-based studies selection process.**

Results and discussion

This section describes the outcomes related to the systematic study RQs discussed above. 51 research studies have been selected to illustrate the outcomes of each RQs. Publication and other selection biases are a potential threat to validity in all SLR and we cannot exclude the possibility that some research studies were missed resulting in reduced precision and the potential for bias. Therefore, we made significant efforts in finding all eligible research articles, and conference proceedings from different well-reputed databases and by contacting experts in the BCT area through social media platforms. We believe that our work provides a significant contribution to the role of blockchain technology in health care.

https://doi.org/10.1371/journal.pone.0266462

Selection results

Our search identified 1177 records, of which 1126 were screened as shown in Fig 7.10. 51 research articles were included in this SLR. The list of selected papers with descriptions of the overall classification results are discussed below.

RQ1: What is the advanced profile used for the employment of blockchain in the healthcare domain?. This SLR addresses the achieved descriptive records about the number of articles that have been published each year, publication source, the average citation received on research papers yearly (see Table 7.3). To complete this SLR, we have examined published surveys, systematic literature review (SLR), systematic reviews (SR), and research papers related to blockchain in healthcare, and published in the field of blockchain from 2018 to 2021. The number of highest citation research articles with the most citations is shown in Table 7.3.

Table 7.3 List of the selected papers with details of QE, publication channel, year, H-index, and citation per year.

Ref	H-index	Publication Channel	Citation per year				
			2018	2019	2020	2021	Total
[43]	68	Neural Computing and Applications	-	-	1	1	2
[44]	29	Journal of healthcare and engineering	-	-	-	1	1
[45]	59	Procedia Computer Science	-	10	22	4	36
[46]	37	Computational and structural biotechnology journal	6	29	26	4	80
[47]	52	Journal of Intelligent & Fuzzy Systems	-	38	21	2	61
[48]	43	Sustainable cities and society	19	67	112	15	242
[49]	76	Business Process Management Journal	-	1	7	2	10
[50]	153	Sensors	-	52	119	21	211
[51]	67	IEEE Internet of Things Journal	-	10	13	1	26
[52]	70	Journal of medical systems	4	14	10	-	30
[53]	70	Journal of medical systems	10	69	135	13	245
[54]	86	IEEE access	21	74	85	14	227
[55]	44	Cognitive Systems Research	15	28	42	6	89
[56]	20	Future Internet	-	2	6	2	12
[57]	105	Future Generation Computer Systems	-	3	18	5	31
[58]	16	Electronics	-	3	38	11	57
[59]	70	Journal of medical systems	3	23	56	1	92
[60]	93	International Journal of Distributed Sensor Networks	1	7	11	1	10
[61]	17	applied sciences	-	1	8	1	10
[62]	70	Journal of medical systems	4	35	42	9	103
[63]	86	IEEE Access	-	5	18	3	28
[64]	108	Oncotarget	18	56	57	11	154

(*Continued*)

Table 7.3 List of the selected papers with details of QE, publication channel, year, H-index, and citation per year. (*Continued*)

Ref	H-index	Publication Channel	Citation per year				
			2018	2019	2020	2021	Total
[65]	68	The Journal of Behavioral Health Services & Research	1	2	28	2	37
[66]	86	IEEE access	-	9	49	4	74
[67]	105	Future generation computer systems	1	17	56	8	88
[68]	60	Wireless Communications and Mobile Computing	-	1	4	1	8
[69]	17	Applied sciences	1	6	39	11	62
[70]	70	Journal of medical systems	-	11	14	3	32
[71]	86	IEEE Access	1	23	39	14	83
[72]	70	Applied Innovation	2	34	46	8	103
[73]	298	Nature communications	1	11	32	4	49
[74]	17	Applied Sciences	-	7	14	2	22
[75]	70	Journal of medical systems	5	41	59	12	136
[28]	37	Computational and structural biotechnology journal	10	78	120	9	237
[76]	127	Journal of medical Internet research	-	3	25	2	31
[77]	70	Journal of medical systems	5	21	26	4	57
[15]	99	International Journal of Medical Informatics	-	-	29	9	46
[78]	93	Computers in Industry	-	-	7	6	13
[79]	86	IEEE Access	-	-	12	4	16
[80]	153	Sensors	-	-	13	3	17
[31]	89	IEEE Transactions on Engineering Management.	-	-	4	2	6
[81]	N/A	International Journal of Environmental Research and Public Health	-	-	-	4	4
[82]	N/A	International Journal of Healthcare Information Systems and Informatics	-	-	1	2	3
[83]	86	IEEE Access	-	-	6	3	10
[84]	-	Electronics	-	-	27	8	38
[71]	86	IEEE Access	1	23	42	12	83

https://doi.org/10.1371/journal.pone.0266462.t001

Fig 7.11 demonstrates the number of articles published each year from 2018–2021. The four obvious outliers are existing from 2018 to 2021. In 2018, 24 articles were published, 16 articles were published in 2019, 9 articles were published in 2020 and 2 articles were published in 2021.

The authors of the reviewed articles were found to be affiliated with institutes located across 17 countries. Five countries, China (number of articles = 12), USA (number of articles = 6), South Korea (number of articles = 4), Brazil (number of articles = 3) and India (number of articles = 3), cumulatively represented 65% of the sample (see Fig 7.12).

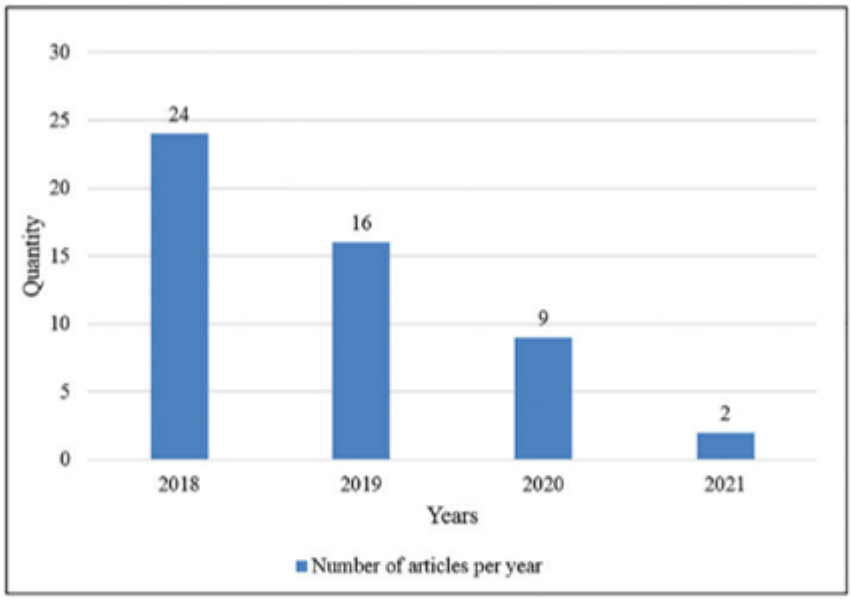


Fig 7.11 Number of articles published per year.

https://doi.org/10.1371/journal.pone.0266462.g002

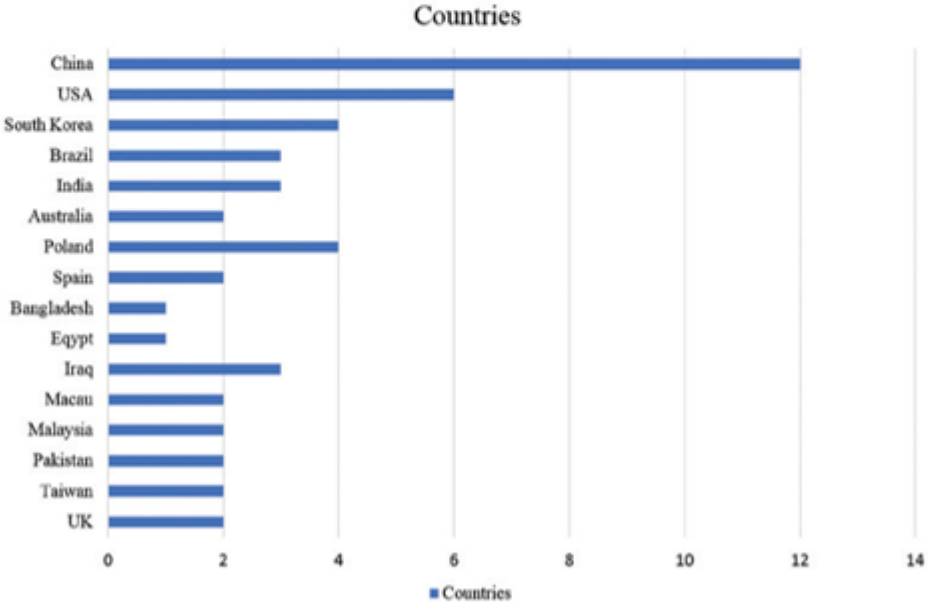


Fig 7.12 Country-wise publications of selected research studies.

https://doi.org/10.1371/journal.pone.0266462.g003

In addition, the analysis of author indexed keywords conducted by using word cloud showed that the main emphasis of article related to "blockchain", "technology", "data", "healthcare", "sharing", and "medical" which are graphically illustrated in Fig 7.13.

Fig 7.13 Constructs key framework.

RQ2: What are the major healthcare domains where blockchain technology has been implemented?. In this part, we will discuss a review of the fundamental principles of blockchain. BCT is used in medicine, especially in managing the information in healthcare that particularly is important in the healthcare area as this technology involves the sensitive data of patients. This sector is important to society because innovations in this field will enhance the quality of people's life. Following this logic, the computation can help to mitigate the effects of certain problems in this field. Informatics, for example, helps in the automation of medical records by ensuring more reliable data sharing, log management, and other applications. One of the first and most popular blockchain applications in healthcare is the exchange of health records. Information related to health is difficult to disclose because it is labeled as confidential information and includes patients' details. Among the key works in the literature that discuss this application of blockchain technology are: [85, 86]. The characteristics of blockchain-based architectures for the sharing of electronic healthcare records can vary. The features of blockchain-based systems for the exchange of electronic healthcare records may vary. One of the most well-known structures in the literature is discussed in the work of Azaria et al. [86]. Several recent papers in the literature have cited this as a framework for the development of other similar architectures. Some of these systems are inspired by Azaria et al. [86] cited in [30, 79, 87, 88]. Voting is a formal statement of an individual's or a group's opinion or choice, whether positive or negative. Traditional voting methods, on the other hand, are centralized and are known to have security and efficiency flaws. The study [89] examines blockchain-based voting systems in depth and categorizes them based on a variety of characteristics e.g., the types of blockchain used, the consensus approaches used, and the scale of participants. Artificial intelligence (AI) is now the core technology for a wide range of applications, from self-driving cars to smart cities. One of the most crucial pillars of social and economic stability is smart healthcare, which is an integral part of smart cities. The research study [90] focused on designing a human-in-the-loop-aided (HitL-aided) scheme to protect patient privacy in smart healthcare. Profile matching technology can facilitate the sharing of medical information across patients by matching similar symptom traits. However, because the symptom attributes are linked to sensitive information about patients, their privacy will be compromised during the IoMT matching process. To accomplish fine-grained profile matching, the study [91] provides a verifiable private set intersection scheme and used a re-encryption technique to preserve patients' privacy. Technologically advanced countries are exploring or implementing smart homes, it is convenient but risky. Most of the existing solutions are generally based on a single-server architecture, which has limitations in terms of privacy, integrity, and confidentiality. While blockchain-based solutions may alleviate some of these problems, they still face some significant obstacles. Lin et al. [92] developed a revolutionary safe mutual authentication method for use in smart homes and other applications. MedRec will be the first sharing architecture to be discussed, which uses a blockchain-based system to store electronic medical records. The MedRec considers resolving issues as data access response time, interoperability, and increased data

quality in healthcare research [86]. It is worth looking into the resources that were used to create MedRec's architecture, since it implements a private P2P network (Permission block chain), as well as using Ethereum's smart contract platform, to make it easier to monitor and track network state transitions. One of the MedRec architecture's hallmarks is that it provides patients with a consulting agency that has records of their healthcare background, enabling them to remain informed about health decisions. Another difference is that they enable the standardization of health data since they are adaptable and provide open data standards in a variety of formats. This architecture takes a novel approach to the use of health data management systems by enhancing security and establishing a common language for data exchange for research purposes [86]. While Azaria et al. [86] also plan to perform experiments and analyses with a diverse community of users. In summary, MedRec is a realistic choice for exchanging healthcare information that can be used to combine patient care, hospital care, and physician care. As a consequence, the reported data can help to minimize discrepancies among different systems of hospitals. As stated by [85], the method introduces the topic of cloud computing, which could help in creating new architectures for sharing healthcare records via blockchain, resulting in safer and more secure healthcare systems for clinical use. The authors propose a cloud-based architecture that uses a blockchain-based data system to connect a network of communication nodes. The paper [85] shows how to handle the exchange of healthcare information using a blockchain architecture, which employs the principles of intelligent contracts and, immutable bookkeeping. The major roles of BCT in sharing health information, remote care with IoT, security, and privacy, and supply chain are depicted in Fig 7.14.

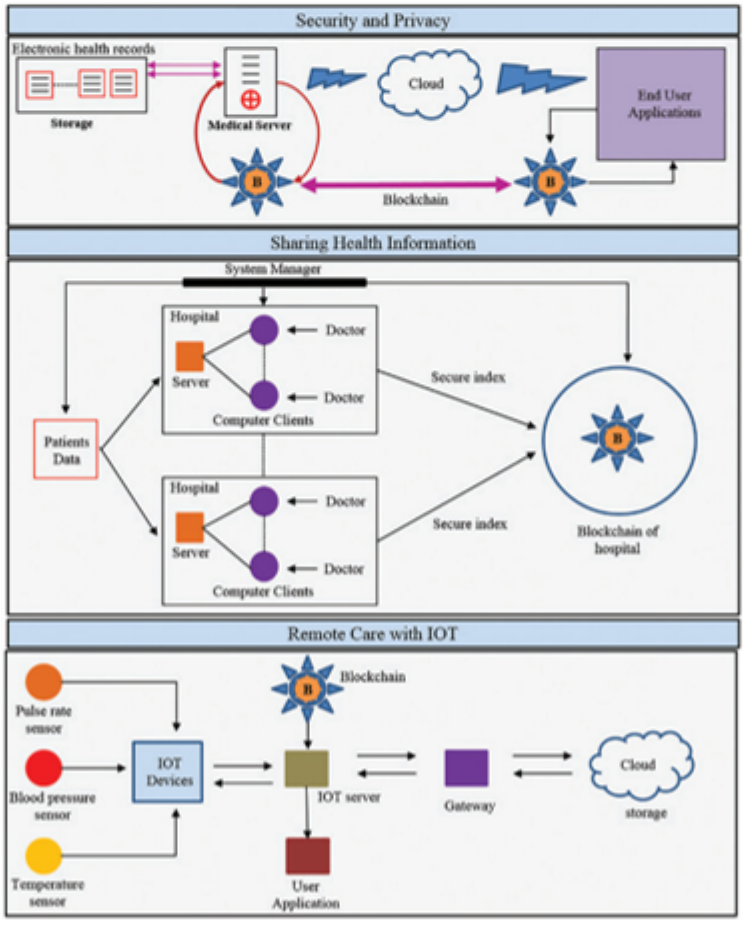


Fig 7.14 Major healthcare domains where BCT has been used.

https://doi.org/10.1371/journal.pone.0266462

The list of blockchain-based healthcare methods is discussed in Table 7.4.

Table 7.4 Blockchain-based healthcare methods.

Ref	Domain	Methods
[86]	Sharing Health Information	MedRec
[93]	Sharing Health Information	MedRec
[94]	Sharing Health Information	Medicalchain
[71]	Remote Care with IoT	Patient-Centric agent (PCA)
[95]	Supply Chain for healthcare	Modum
[96]	Security and Privacy	Decentralized sharing of health records (DSHR)

https://doi.org/10.1371/journal.pone.0266462.t002

RQ3: What are the existing problems and constraints raised by the previous studies in the healthcare field using blockchain technology?. Even though blockchain is a multidisciplinary concept with challenges and limitations, it can be applied to a variety of areas [88]. Researchers in this field are working to overcome or mitigate the negative effects of these factors. The following are some of the problems (i.e. technological challenges) that blockchain technology faces when used in healthcare [22, 88, 97, 98].

1) Throughput. If the number of transactions and nodes in the network grows, more checks will need to be performed, possibly causing a network bottleneck. When dealing with healthcare systems, high throughput is a challenge because unless there is fast access, it might adversely affect a diagnosis which could save someone's life [64], correspondingly recognizing that the suggested framework focuses on identifying inconsistencies will possibly not perform well when datasets are unlabeled. Issues including specifications for continuous updates by the used system [59], keyword set size [75], network set-up and disk space needed based on the blockchain type, such as Ethereum software, employed to the framework can all affect a framework's scalability and performance quality [25, 99]. Similarly [48], suggests that integrating certain features into established systems, e.g. making global smart deals, can offset higher performance-related costs. Furthermore, a small number of studies have suggested that performance-related problems can be connected to the node management in a suggested system.

2) Latency. Validating a block takes about 10 minutes; this can be harmful to system security services since successful attacks may occur during that time. Healthcare networks are complex and should be accessed at all times, as any delay may negatively affect the analysis of an exam.

3) Security. When a party has control of 51 percent of the voting power, this can adversely affect the computing power of the network. This is a serious issue that needs to be addressed because a harmed healthcare system will lead to healthcare organizations losing their reputation.

4) Resource Consumption. Since the mining process consumes a lot of energy, using this technology could result in a significant loss of resources. Since multiple devices are required to track patients in a healthcare setting, energy costs are high; however, the use of blockchain may result in high computing and energy costs. Managing these expenses is a challenge for businesses.

5) Usability. Since these systems are so complicated, usability is a challenge as well to deal with. Additionally, an API must be developed (Application Programming Interface) Users would enjoy the user-friendly features. Since not all health practitioners have the same level of education, As IT professionals, we should be able to use frameworks that are easy and effective.

6) Centralization. Even though blockchain has a decentralized design, certain implementations tend to concentrate the miner, which decreases network stability. Because this central node is insecure and may be hacked, hostile attackers can get access to the data it holds [25].

7) Privacy. It is common to suppose that the Bitcoin framework allows blockchain to make sure the privacy of its nodes. The results of [25], on the other hand, contradict this assumption. Furthermore, strategies to provide this functionality to blockchain-based systems are needed [25]. Due to privacy laws and regulations, blockchain-based systems have to conform to the General Data Protection Regulation (GDPR). Our research also indicates those users' reservations about the safe and ethical utilization of data could be a major barrier to blockchain adoption in healthcare systems. The existing issues are primarily associated with blockchain technology's technological limitations, such as the protection of individual nodes [53], the degree of safety permitted through cryptographic elements implemented with the system [70], besides the preservation of confidential data whereas requesters complete their computations [100]. However, certain research has drawn attention to more socially relevant issues regarding sharing of public data [73] and users' confidence in governments [49, 74]. Such issues may also be linked to the suggested framework protection from the perspective of users for example users' management and misuse of permitted personal keys/codes [46].

8) Constraint. Prior studies have identified constraints, which can be divided into four categories. These dimensions mean that such constraints extend beyond technical boundaries (costs of designing, implementing blockchain systems, data analysis for system assessment, and framework constituent elements) to include certain social facets also such as trust in the administration, infrastructure of technology in a country.

9) Costs. This set of constraints is specifically concerned with the time, capital, and economic expenses of putting a system of blockchain into action. For example [50], discuss resource constraints in IoT, while [28] discuss the costs of arranging dispersed app in the deployment of blockchain. Additional expenses that have been established as constraints and limitations in previous research include the linear increase in protocol costs based on the characteristics and attributes of the entities involved, such as patients [54], increased operational overhead for the patient, and access latency for the requester [69], the exchange and implementation costs depend upon inconstant inputs in size and length of a string [67]. The issues related to time are further listed as one of the limitations, i.e. the spent time in finding smart contracts globally [48], increased time consumption [57], transmission timing [53], the time needed for the data receiver to seek the required data in shared storage [68], and higher overall execution time [21].

RQ4: What are the potential healthcare avenues that would benefit from blockchain technology implementation?. Blockchain consists of a sequence of blocks connected with cryptographic techniques. The immutability of this is one of the most attractive characteristics to many industries. The data that is added to the blockchain is irreversible, consequently, allowing for the creation of a consensus-based, verifiable, and accurate data ledger. That creates blockchain especially well-suited to tasks wherever integrity of data is critical; ProvChain [101], an infrastructure based on this technology in giving chain-of-custody to the database, is a functional example of this immutability. There are many blockchain implementations, including Bitcoin, a cryptocurrency token based on the blockchain; and Ethereum, a cryptocurrency token based on blockchain. Ethereum [102], a blockchain ledger with Turing-complete computer-generated device that allows smart contracts to implement code on this; and JP Morgan's Juno [103], an Ethereum fork that uses the particular consensus mechanism called Quorum, along with several other blockchain implementations. The execution of blockchain varies due to ways in their consensus approaches. Bitcoin, for example, employs the HashCash [104] Proof-Of-Work algorithm, which is a deliberately slow system intended to avoid denial-of-service attacks. As a vote against the blockchain's agreement, every Bitcoin miner authenticates this blockchain system by conducting that algorithm. Ethereum includes Ethash, which is an algorithm called Proof-of-work based on the Dagger-Hashimoto algorithm, as described in the Ethereum Yellow Paper [103].

However, shortly Ethereum is likely to advance in an algorithm named Casper. It will consider the excess requirement of energy in Proof-of-work [105]. The implementation of smart contracts separates Ethereum from Bitcoin. The smart contract is one of the snippets of code that run on each blockchain node. These are self-executing contracts in which all members of the blockchain are bound by the agreement. In the same way, as a standard contract does, they influence advantages, responsibilities, also punishments related to contract-related conduct. It could be utilized to model the HIPAA healthcare personal health information (PHI) workflow to satisfy audit and regulatory standards, likewise, done inside Patientory since they resemble conventional paper contracts and rules [106]. A new type of blockchain trust model, trust in the consortium, is also emerging. Microsoft recently released the Coco framework, which enables the creation of blockchain-agnostic consortiums [107]. Above mentioned models are based on a pre-defined group of trustworthy parties. It can be among various clinics or in the UK, NHS Trusts, third parties, and manufacturers of devices. By implementing smart contracts only on the hardware of trusted partners, without requiring miners, a consensus can be generated. It turned out in remarkably improved results, through a Coco-optimized blockchain case capable of processing 1600 transactions in a second, taking the blockchain system very close to the major payment processors. Coco also supports a variety of trusted execution environments, including Windows Virtual Secure Mode, Arm Trust Zone, and Intel Guard Extensions to name a few.

1) Clinical trials. Managing trial subject consent and clinical trials itself is an area in which blockchain can potentially improve the accountability, audit ability, and transparency of researchers and practitioners in the medical field. By keeping the unchangeable log of a patient's approval, officials could control the standard of clinical trials easily, making sure it complies with informed consent regulations of the country. It is especially important because a forged informed consent form is one of the common types of clinical fraud. It involves falsifying patient consent and editing records, implying that authentication of trial subjects is essential for avoiding it. That kind of setup may be improved by implementing a smart contract system that stops clinicians to use the data of patients unless a key is issued by the end of an auditable process of smart contract that requires permission in each step in the trial, as proposed by Benchoufi, Porcher, and Ravaud. This procedure should also allow the patient's consent to be revoked. Executing the clinical trial of blockchain consent log provides the subjects with data ownership while also having a trail of audit for regulators, medical professionals, and researchers. The role of BCT in clinical trials is graphically represented in Fig 7.15.

Fig 7.15 Role of blockchain in clinical trials.

https://doi.org/10.1371/journal.pone.0266462

2) Sharing the data. Sharing information is regarded as the most significant opportunity for improvement in healthcare; however, it too poses a significant challenge in privacy. Sharing the data using BCT is presented in Fig 7.16. Powles and Hodson [108] use DeepMind's case study teamwork with the Royal Free London NHS Foundation Trust to address the need for transparency in how patient data is being shared with third parties. Regardless of the good impact on diagnosis/treatment of patients by the product suite of Google, one of the significant issues addressed in the previous case study was a lack of patient consent. On the other hand, Sleep Apnea American Association, and IBM [109] were collaborating to solve major healthcare challenges to examine sleep apnea (with IBM's Watson supercomputer at home) in thousands of Americans, with informed and clear patient consent. That was critical to implementing the national standard for interoperability in the healthcare system of IT. Which was emphasized by Wachter and Hafter through a white paper in UK NHS in comparison to the US healthcare sector that emphasized the significance of interoperability in permitting patient Electronic Health Records (EHRs) over various clinics, such as various trusts that do not maintain a separate system to get access on these records built by different vendors.

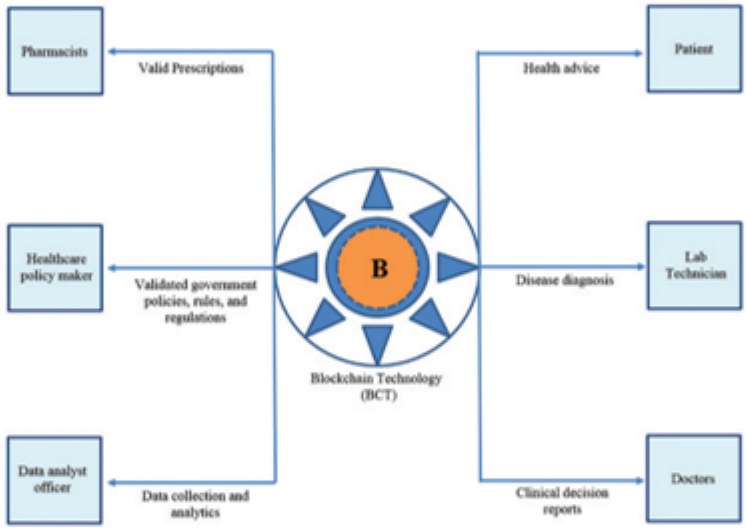


Fig 7.16 Sharing health information using BCT.

In a report of Harland Simon on a project justifying RFID tagging in NHS Cambridge shire, about 15% loss of assets annually, resulting in a substantial cost to repurchase the items that hospitals previously have. Furthermore, as per GE Healthcare [110] report, nurses spend an average of 21 minutes per shift searching for misplaced devices as stated by, defining any device under $5000 as consumable and to again purchase if any device is lost, suggesting significant cost in the sector. Published by Harland Simon, another study reported that [111], by adopting radio frequency identification (RFID) standards for tracking of medical devices, NHS Forth Valley in Scotland had saved nearly £400,000 in cost avoidance by not having to buy the important devices which would have been lost by the medical system. Tracking of Drugs has been a completely different issue than tracking of devices because a major concern is counterfeits of drugs here. According to WHO's report, In the US up to 10% supply of Pharmaceutical products is counterfeit. In the United States, the Food and Drug Administration (FDA) recently approved the utilizing the RFID to track pharmaceuticals between the supply side to the patient. It enables the whole sequence kept supervised, to make sure that pharmaceuticals were purchased from a legitimate source. Pfizer was the first pharmaceutical company to use RFID "e-pedigree" to ensure that patients and doctors could trust the source and capabilities of their flagship medicine, Viagra, after identifying it as one of their most counterfeited drugs. Because of the

use of low-cost passive RFID tags and barcodes, the system enabled the pharmacists and wholesalers to check the authenticity of their Viagra through a simple RFID scanner at a low rate than Pfizer.

3) Records of patients. Blockchain is having the potential to significantly disrupt health services and place data in the patient's hand. The specific intriguing steps are in MedRec [86], which provides doctors and patients with an immutable log of a health record as shown in Fig 7.17. That has a different approach to incentivize miners by providing access to anonymized data about health in exchange for network maintenance. MedRec maps Patient-Provider Relationships (PPRs) using Smart Contracts when the contract displays a reference list having relationships between nodes on the Blockchain-system. This too places PPRs in the patient's hand, empowering them in accepting, rejecting, or modifying relations with health service providers for example doctors, insurers, and hospitals, etc. Blockchain-system allows for interoperability in the health system by providing a decentralized ledger of accepted facts in healthcare records to which all health service providers are having access. It implies that while user interfaces may differ, the central ledger would be the same across all service suppliers. A challenge that exists relates to the current state of health records across providers, which contain significant amounts of the same information under different identifiers that may not be linked. This causes replication, and as the blockchain system increases in size, it is reduced in performance. The level of data duplication in all records will necessitate replication to maintain a reasonably performant system with unique, anonymized identifiers to identify the patient in all kinds of service. Adopting the blockchain health record is a business challenge in and of itself. The important thing is that medical records will not start from zero because they would have to replace the current setup, and that is challenging. Furthermore, the sheer volume of data generated in the healthcare sector is ever-growing, with Kaiser Permanente estimated to have between 26 and 44 petabytes of data on its 9 million members from EHRs and other medical data in 2014. The data volume which is logged and referenced would mainly exacerbate the scalability issue.

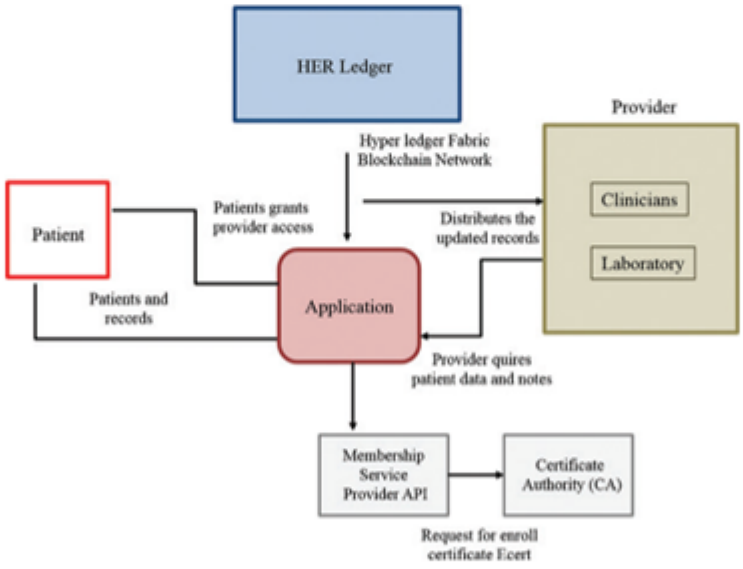


Fig 7.17 Records of patients.

4) Drug tracking. Another opportunity is tracking the drugs using a blockchain system as shown in Fig 7.18, which takes advantage of its immutability in the development of tracking and a chain of custody from manufacturer to patient. Chronicled is a technology startup company that is working on its product, Discover, that develops a chain of custody model that shows the manufacturing place of a drug, the places it had been since then, and when it was disbursed to patients, hence reducing the pharmaceutical theft and

fraud. That enables the health professionals in meeting existing standards of the pharmaceutical supply chain, along with focusing on interoperability among healthcare professionals. The Counterfeit Medicines Project has been launched by Hyperledger, the Open-Source Blockchain Working Group, to address the issue of counterfeits of medicines. The origins of counterfeit medicines would have been tracked and thus eliminated from the chain of supply. One benefit of tracking drugs by blockchain-system over conventional methods is the inherent decentralization of trust and authority in the technology's principles; whereas chief authorities could have bribed or faked, it is much more difficult to bribe a consensus of those on the blockchain. As a result, an existing standard in pharmaceuticals tracking in industry, ePedigree, which already employs RFID and a traditional database, is transitioning to its blockchain application. If medicines/drugs could be tracked and developed at the point of manufacture using blockchain's inherent anti-tampering capabilities, that will remove the counterfeited pharma products engaging in the supply chain.

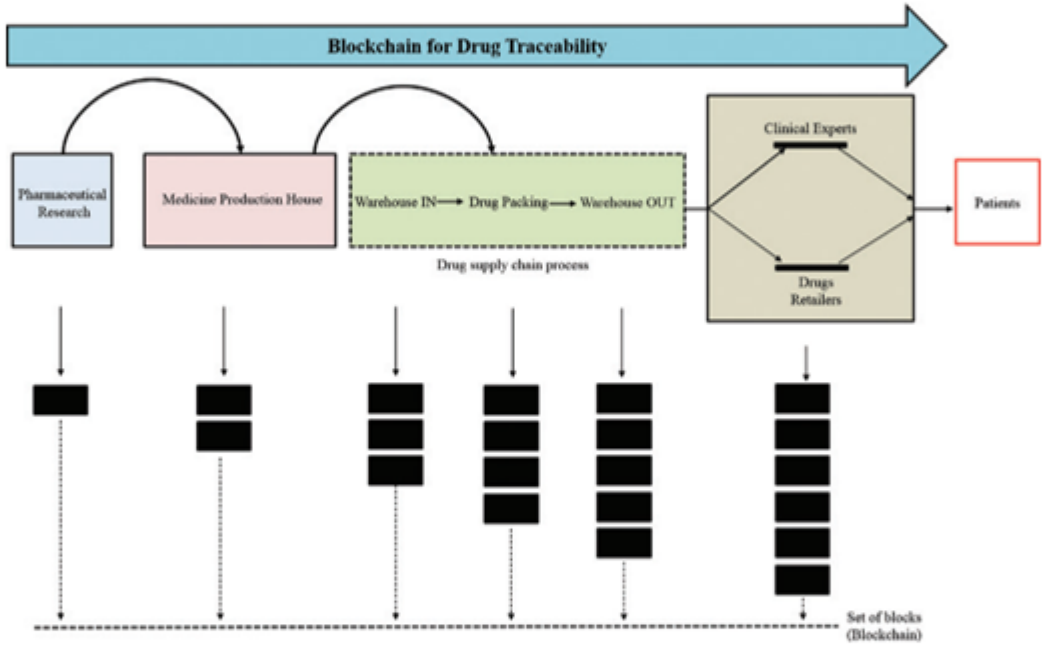


Fig 7.18 Drug tracking.

5) Device tracking. Tracking of medical devices is one aspect of Block chain in disrupting healthcare, from manufacturing to decommissioning. The monetary benefit generated through tracking of assets is clear; NHS East Kent Hospital discovered 98 infusion pumps they had no idea they still owned across three sites as concluded through the case study of Harland Simon [111] in which active RFID trackers were implanted. Because of this single case study, they saved $147,000 at $1,500 per person. The use of blockchain in conjunction with this technology allows for an immutable ledger that not only shows the current location of the device, also the location of the lifecycle, along with the serial number, distributors, and the manufacturer linked with the device, assisting with regulatory compliance. Deloitte identified the competency among the potential game-changers for blockchain in the domain of healthcare in a white paper. According to an IBM study, 60 percent of government stakeholders in healthcare believed the integration of medical devices and asset management as the most likely area of disturbance in industry. The blockchain-based system has various advantages above conventional products of location tracking. This immutability and tamper-proof properties of the Blockchain are the most obvious. This prevents a malicious user from changing or deleting a device's location history. This is especially important given that theft of devices and shrinkage has been

a major issue in the United States and the United Kingdom. This immutability, in addition to preventing traditional theft, also protect devices from being lost and reordered, that have incurred high cost in terms of both actual equipment cost and care provided. The setup must not add significantly to the workload of staff, nurses, or workers because that requires tapping the device only using a mobile phone and further entering the device location. Whereas the use of blockchain on the Internet of Things (IoT) is still in its early stages, Huh states a method to communicate the devices using an Ethereum blockchain and public key system of RSA. Likewise, the device while it stores on blockchain its public key also stores the associated private key on the device.

Discussion

This study aimed to conduct an organized analysis of previous literature about the employment blockchain in healthcare for a better understanding of their current and probable state. The four key research problems are defined for this reason. RQ1 was presented with summing up top writers, publishers, publication houses, and designs of publication patterns of this subject. Furthermore, it included an existing outline of research about employing blockchain in the healthcare space. The comprehensive description of the reviewed articles is discussed in Table 7.5. RQ2 was designed to help researchers better understand how blockchain can be used, it is responded by defining particular themes and sub-themes that reflect key aspects in employing blockchain in this sector. RQ3 further discussed its shortcomings and obstacles that previous researchers had encountered. We were able to recognize the research gap in the existing literature and responded to this question by summarizing its main research themes and existing limitations. RQ4 concentrated on the key aspects where future investigation can provide valuable insight. The fourth research question is addressed by combining findings through emerging differences, shortcomings, and previously proposed guidelines.

Table 7.5 Comprehensive description of reviewed studies.

Ref	Methods	Description	Channel	Proposed / Implemented	Components
[45]	Public blockchain	Details about the algorithm are provided.	multi-tier, public	Implemented	PBEDA, ECDH, MVP, and ECDSA
[46]	MAM is used for real-time broadcasting activity via wearable's	Algorithm details are provided.	IOTA	Implemented	Masked Authenticated Messaging for Tangle, Merkle hash technique, Merkle signature scheme, One-time signature
[47]	Computing Edge	There was little information available on the outcomes and efficiency.	The channel was Ethereum	Proposed	WSN and Wireless sensor network controller v.2, computing Edge
[48]	Not Available	MedRec was used to do a cost analysis, but no algorithm or data were given.	The channel was Ethereum	Proposed	Ethereum Go-client, QuorumChain algorithm, classification, smart contracts, cipher manager, proxy re-encryption

https://doi.org/10.1371/journal.pone.0266462

Ref	Methods	Description	Channel	Proposed / Implemented	Components
[49]	UID system	Details about the algorithm are not provided.	Ethereum channel	Proposed	UID system, analytics of Big Data, history & registration contract
[50]	Not Available	Details of the algorithm are provided, along with security margins that have been verified against defined criteria.	The channel was Bitcoin	Proposed	Internet of Things, Merkle tree, Diffie-Hellman key exchange, and digital ring signature
[51]	Not Available	Details about the algorithm are provided.	The channel was Consortium	Implemented	Practical Byzantine fault tolerance consensus mechanism,
[52]	PSO	Performance results from a static study based on specific assessment criteria	N/A	Implemented	ADB, boosting for ML, reverse engineering, feature selection & extraction
[53]	Not Available	Details about the algorithm are not provided. no experiments.	Federated blockchain	Proposed	IoT sensors WBANs, oracle
[54]	Proposed signature scheme based on Attribute	Details of mathematical and computational notations given for execution of scheme, performance, and security evaluation	Not Available	Implemented	Diffie-Hellman computational bilinear, MA-ABS scheme
[55]	With the support of a genetic algorithm Proposed access control method based on blockchain	In terms of input and output strings, block creation, time processing, detailed simulation results are provided.	N/A	Implemented	Genetic algorithm, hash key cryptography, MD5 strings, discrete wavelet transforms.
[56]	The proposed Blockchain-based eHealth Integrity Model uses a design-science methodology.	The integrity-verification algorithm's details and test results have been released.	Permissioned blockchain	Implemented	Byzantine Fault in Practice Algorithm for tolerant consensus

(*Continued*)

Table 7.5 Comprehensive description of reviewed studies. (*Continued*)

Ref	Methods	Description	Channel	Proposed / Implemented	Components
[57]	(HAR) Methods for recognizing human activity based on a uni model	Some mathematical notations and comprehensive findings for performance reassessment studies on three datasets have been given.	N/A	Implemented	For HAR (ECOC) framework, (SVM), Multi-class cooperative categorization technique, fog computing
[33]	Not Available	For performance evaluation, precise mathematical and algorithmic notations are supplied, as well as the outcomes of experiments.	N/A	Implemented	The Merkle tree with order-preserving encryption
[59]	Not Available	The applicability of blockchain in healthcare is discussed using a concept-based approach.	N/A	Proposed	Environment based on the blockchain
[60]	Preprocessing of images	Details on the experimental results, classification training, and testing outcomes were given.	Bitcoin	Implemented	HOG, LBP, SVM, RFT, DNN
[61]	Pearson's correlation, compression ratio technique	The compression ratio and stability performance testing results were given.	N/A	Implemented	(BAQALC) proposed, (NGS), (SRA), LZW modification
[62]	Methods of authentication	Details on the algorithm and how it performs in areas of operational costs were given.	Ethereum	Implemented	primitiveness verification (PV), preservation Submission
[63]	Method of the primal-dual Varangian	The performance study results of a strategy for achieving Stackelberg equilibrium were disclosed.	N/A	Implemented	three-layer (Hierarchical architecture), edge computing

https://doi.org/10.1371/journal.pone.0266462

Ref	Methods	Description	Channel	Proposed / Implemented	Components
[64]	Machine Learning(ML)	In place of simulation, experiment-based assessment, the proposed architecture is presented through workflow examples.	Exonum	Proposed	Inbreeding coefficient, DNN predictor, data temporal value, LifePound (utility crypto token)
[65]	Not Available	There are just a few algorithmic notations, but comprehensive findings for Apache JMeter performance evaluation are supplied.	N/A	Implemented	Merkle tree, timestamped algorithm, Keyless signature infrastructure
[66]	Not Available	Code scripts, access control methods, and performance assessment findings for access control and network overheads are all shared.	Permissioned blockchain on Ethereum	Implemented	(IPFS), (ABE), mobile cloud
[67]	Not Available	For performance evaluation on many parameters, details of algorithmic notations are given together with findings.	Permissioned blockchain on Ethereum	Implemented	MediBChain protocol, (ECC)
[68]	Not Available	For performance parameters, there are just a few algorithmic notations, but extensive simulation results are supplied	N/A	Implemented	The one-time transaction, Ring signature algorithm, stealth address, Cryptonote protocol
[69]	Not Available	For numerous parameters, with the discussion of outcomes, there are few algorithmic notations from experiment-based and theoretical mathematical results	Permissioned blockchain	Implemented	consensus algorithm BFT smart, ECC, MedChain, modified digest algorithm

(*Continued*)

Table 7.5 Comprehensive description of reviewed studies. (*Continued*)

Ref	Methods	Description	Channel	Proposed / Implemented	Components
[70]	Not Available	For simulated security analysis and performance evaluation, few explained algorithmic notation results are used.	Blockchain technology Hyper ledger	Implemented	SIFF
[71]	Generating Sessional symmetric key	An in-depth look into simulated performance and security assessments.	Custom bitcoin and Ethereum	Implemented	PUA, Trei tree, mutual authentication protocol, HMAC
[72]	Not Available	There are no experimental or simulation test results, but algorithmic notations and the syntax are described with analysis of specified security.	Consortium blockchain	Implemented	ABE and IBE, proposed d identity-based combine attribute, identity-based signature, signature, and encryption
[73]	Not Available	There was only a brief discussion of the outcomes of simulated blockchain-based clinical trials. There is no explanation for the algorithm or syntax.	N/A	Implemented	artificial healthcare, Parallel healthcare system, IVRS, parallel execution
[74]	Not Available	There is only a brief explanation of the method and the findings of the prototype implementation, which is mostly theoretical.	Consortium blockchain technology	Proposed	proposed proof of familiarity(PoF), API
[75]	Proposed protocol privacy-preserving and Secure PHI sharing BSPP	Algorithm notations, system architecture, and protocol implementation, and performance assessment findings are all detailed.	Private, Consortium	Implemented	Bilinear maps, Consensus mechanism

https://doi.org/10.1371/journal.pone.0266462

Ref	Methods	Description	Channel	Proposed / Implemented	Components
[28]	Case study involves	Through a case study, a detailed explanation of the proposed architecture and process is provided. There are no experiments or discussions on the algorithm or syntax.	Ethereum technology	Proposed	Oauth, Public-key cryptography—sign then encrypt mechanism, Solidity smart contract, FHIR,
[76]	Not Available	There is not much explanation of the algorithm, but there are a lot of details about the experiment that was done to verify the system's feasibility.	IOTA Tangle (DLT)	Implemented	(MAM) Masked authenticated messaging, GPS, IoT integration
[77]	Not Available	Algorithm creation and performance assessment for processing time and transaction verification are discussed.	Ethereum technology	Implemented	MIStore, PBFT

https://doi.org/10.1371/journal.pone.0266462.t003

Conceptual evolution

According to the findings of the study, research in healthcare's blockchain was largely focused on enhancing more creating new ideas and concepts that help researchers to derive multi-domain [59] also practicable blockchain in healthcare implementations. The viability of employment [99] is being established and evaluated across three sub-themes of research: design creation, applications on benefit-based, and developing predictive competencies.

Concept development

The findings of the analysis indicate that new proofs and algorithms have received a lot of attention, such as proof of data primitiveness [62, 112], proof of familiarity [74], and simpler workload for proof of work; [54]. Studies have also focused on testing new variables and components in architecture systems, as well as improving frameworks that enable blockchain execution by including them. Consider cryptosystems based on attribute [72], approach the Stackelberg game [63], sibling intractable functions [70], and homomorphic computations for more efficient frameworks [77]. Further [113], suggested a new scheme (BBDS) based on blockchain to protect data transactions and maintain privacy [57]. Used fog computing estimation efficiency as well as reliable models for human pursuit acknowledgment to support remote e-health controlling [51].

To eliminate a single point of failure, they are focused on incorporating several time sources into their technique. Finally, this research has centered on how to boost the efficiency of already established algorithms and structures based on the blockchain.

Benefit-based application

Blockchain has been used in healthcare research to extract concrete benefits by identifying and testing new technology avenues. It involves work in upgrading the technical advantages of employing blockchains, such as advanced image processing [60], effective behavior recognition [57], and Internet-of-Things synchronization (IoT) devices [51]. Furthermore, the majority of studies in this category have centered on the use of blockchain technology to establish specific benefits of healthcare, e.g. mutual decision making in the medical field [74]. Blockchain adoption, for example, is being suggested to have positive implications while managing clinical trials [73], DNA data transmission [61], preventive healthcare, biomarker growth, and discovery of drugs are all examples of remote patient monitoring [50, 53, 64].

Advancing decentralization

Existing research is also considering promoting key advantages of blockchain technology throughout healthcare environments for encouraging justice and also efficient decentralization [76, 114]. For instance [63], produced an efficient framework for promoting maximization of revenue maximization along with fair decentralized trade, considering that [48] stated the need for trade-offs for mining benefits. Researchers in previous literature already described blockchain's prospects in developing transparency in exchanging the data [56], such as utilizing upright client roles [21, 30]. By doing this, we can say that previous research about the increasing use of blockchain-based technology in the medical space is focused on spreading decentralization along with its related advantages.

Advancement in technology

Current studies have contributed substantial progress in terms of advancement and refinement of blockchain for the development of targeted deployment, particularly in the healthcare space. We suggested previous research that is classified in this theme be directed regarding the three main topical issues based on our review:

Developing intelligent healthcare ecosystems

The introduction of blockchain technology programs into healthcare environments has piqued the interest of some academics [45]. Such integrations can pave the way for the development of intelligent healthcare systems [100]. For example [47], argues that blockchain adoption will aid in the development of a more efficient e-health ecosystem. Prior research has also suggested frameworks for developing blockchain-based e-health [56] and telemedical information systems [33], which could help healthcare providers, expand the scope of their services in the future.

Improvements to the blockchain architecture on a technical level

The majority of this field's study has concentrated on improving the efficiency of architectures and developed systems by technological improvements for example utilizing smaller data block sizes [74] and reducing transaction propagation delay [68, 74]. Some attention has also been given to problems that have previously been described as possible roadblocks to the successful implementation of blockchain architectures. Memory and CPU specifications [77], as well as accurate node recognition, are among the problems considered in research, grouped under this theme [71]. The efficacy of potential solutions to the above problems has been

illustrated in several cases by network and algorithm comparison studies [21, 30, 74]. However, we believe that this theme will continue to progress in the future, necessitating a parallel emphasis on comparative analyses to determine the most powerful networks and algorithms.

Building predictive capabilities

A similar pattern can be seen in blockchain technology's use in healthcare as it enters the fourth step in development with the rising integration of AI [26]. IoT [50], sensors [47], wireless body area networks [53, 64], big data [49], edge computing [76], and cloud technology have all recently been incorporated through blockchain-based device architecture [59]. Researchers are using such technologies to help them develop systems based on blockchain having predictive capacities for enhancing medical information systems and diagnostics [61, 75]. Prescription fraud avoidance [47], verifiable data generation [77], and automatic claim resolution [47] have all been investigated previously using such frameworks [72]. Furthermore, studies have centered on using blockchain-based technology in supporting providers of health care services with other tasks, e.g data collection on population-level [46] and user identity description [28].

Enhancement of efficiency

Several researchers in previous literature have attempted to determine how blockchain-based implementation would improve the efficiency of healthcare processes [34, 59, 79, 82]. According to this study, scholars' attention has been drawn to two facets of performance improvement: structures and processes.

Process

Prior research has focused on improving the efficiency of technological aspects of the processes that are needed to run a blockchain-based healthcare system. Prior research has also focused on developing systems for timely alerts [72] and adverse event reporting [73]. Some research has centered on increasing the computational processes efficiency [57] and thorough evaluation of suggested architectures [60] to ensure that its architecture gives far efficient processing as compared to conventional architectures [71]. Furthermore, studies have proposed changes in blockchain systems to resolve the alleged risks related to time management, data management, and managing related costs. For example, reviewed research has established mechanisms for reducing the cost of implementation after setting up initially [74], lowering storage costs [65], and making maintenance and storing files of any size is easier [62, 112].

System

According to our analysis of the current literature, several steps have been applied to enhance the blockchain-based healthcare framework holistically. For example, research has focused on improving system interoperability [27, 49], managing inter-institutional access rights [63, 99], and data management [28, 46, 63, 64, 99]. Enhancing machine scalability [113] and efficiency have also received scholarly attention [60, 65]. Researchers have concentrated their efforts on designing integrated service-oriented architectures [56] and enhancing the generalizability and flexibility of blockchain systems that have been implemented [21, 27, 30].

Management of data

According to the results, we can conclude that managing the data and medical records is getting more scholarly attention. Existing research has endorsed the blockchain's utilization for medical data management [27, 48, 54, 55, 69, 70, 99, 115]. Furthermore, by integrating heterogeneous forms of data [27, 69, 100], blockchain

can aid in the development of an application system of information to manage such PHRs [59, 64]. We define three main aspects of current research in this field based on the SLR.

Data privacy

Previous research about data security implications of this technology in medical care has focused on handling the privacy of data by maintaining permitted access to the data. According to the study, access control management [76] has gotten a lot of attention [45, 46]. Because of the requirement of protecting the privacy of confidential data by greater transparency, access control, and immutability, this problem is particularly important in healthcare [55]. Prior research has developed a framework based on blockchain to guarantee the delivery of effective services [115], user-centric [114], and access to patient PHRs and other medical data that is safe and encrypted in response to this vital need e.g. [45, 50, 54].

Data protection

Another main concern discussed in studies on the blockchain aspects of data management in healthcare is the avoidance of unauthorized access and the preservation of data confidentiality to ensure data safety. The majority of the studies that were examined focused on preventing unauthorized access [66] and preventing eavesdropping [71]. Several methods have been proposed to achieve this aim, including efficient authentication [65], biometric authentication [49], user verification [55], and the use of dual signatures [63].

Data handling

Prior research has addressed the need for legally and legally compliant collection, sharing, and controlling of healthcare data to some extent. According to our findings, few research studies have recognized the importance of monitoring enforcement [100], let alone the criteria and targets for compliance [63, 67, 75]. However, the importance of data integrity has received a lot of attention [56, 69, 70]. Prior research has looked at issues like authentic data mobilization [46, 77], double storage expenditure [68], and eternal data protection [68, 112, 116, 117]. Along with the growing inter-institutional adoption of blockchain, researchers have transformed their attention to the issue of storing and maintaining sensitive data [67] from a variety of sources, including medical devices [52] and health insurance [77]. A few studies have concentrated on the assistance of cross-institutional sharing of data [67], as well as changes in data sharing quality and flexibility [69]. Additionally, previous studies have addressed the need for data processing improvements (e.g. [77]). Some steps for inducing these changes have been suggested in the reviewed studies, such as the successful incorporation of diverse data from various sources of data [99] and the integration of smart contracts [48]. These themes specify that past studies in that area have focused on (i) improving technological features, (ii) managing medical databases, also (iii) identifying unique capabilities in the medical field, where blockchain could make remarkable contributions. Based on emerging trends, it can be said that scholarship in this field is still transforming, with existing facets of healthcare being recognized as possible recipients of using blockchain as a result of technological advancements.

Research framework for future synthesis

This analysis and review helped in the framework development which was created with the research gaps identified in existing recommendations suggested in previous research. The research model includes 5components that would aid in the development of the healthcare ecosystem based on blockchain, for future research. The research framework for the BCT-based healthcare system is depicted in Fig 7.19.

https://doi.org/10.1371/journal.pone.0266462

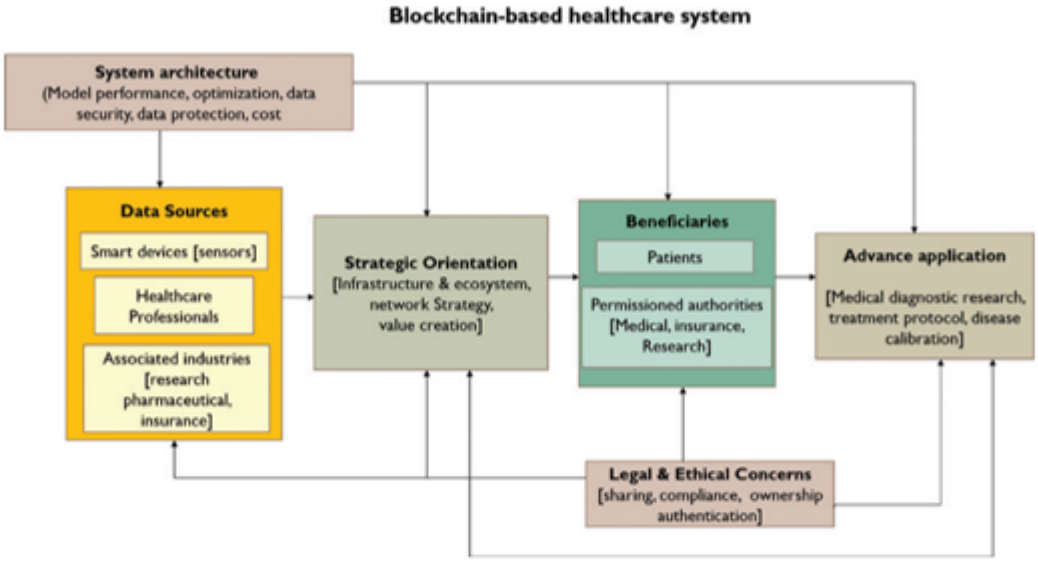


Fig 7.19 Research framework.

https://doi.org/10.1371/journal.pone.0266462.g010

Data sources. Personal and medical health records are created and managed by the patients using mobile devices, healthcare service suppliers, and pharmaceutics is one of these associated industries., research and insurance [32, 47, 70]. These serve as the foundation of the blockchain architecture and need management by legitimate and regulatory rules. This technology might aid in the development of authorized databases that information can be retrieved by inter-institutional authorities in collaboration with the required agencies to aid in patient treatment and medical decision-making [21, 30]. Because of the incorporation of newer technologies, such as smart patient tracking devices [53], to increase the comprehensiveness of medical databases, future research should concentrate on handling those data sources.

System architecture. With advances in blockchain technology, the system's architecture will undergo important changes and refinement in terms of the components incorporated into the system of blockchain. Such as using Permissioned consortium blockchain [28] or platforms other than Ethereum [68, 77], could improve current blockchain deployment architectures in the healthcare ecosystem. Also, future research should concentrate on creating techniques for managing system architectures that have been established, particularly the challenging circumstances that can have an impact on the performance and efficiency, for instance, node management [74] and techniques of key distribution [67].

Blockchain technology strategic implementation. In the case of increasing integration of information and communication technology and blockchain over healthcare ecosystems, researchers should have focused on the elements which could impede and assist in the widespread application of blockchain technology. According to our findings, we believe that organizations should think about whether or not, by identifying the key issues throughout this study, blockchain technology might prove a potential source of creating or enhancing value. Strategic problems like resource constraints [50] and technical problems like performance uncertainty [56] and system requirements are these examples [99]. Considering these problems might help scholars to develop blockchain architectures that can offer better functional utility and productivity in terms of resource and output management. This can also guide health system administrators and personnel to adopt a holistic and strategic approach to the potential inclusion of blockchain as an essential component of a company's value chain.

Beneficiaries. Databases built on the blockchain can provide trustworthy information to particular beneficiaries in the sector of healthcare, such as patients who keep ownership of their information. Authorities including doctors, pharmacists, medical researchers, and insurance firms are also beneficiaries. Patients may authorize them to use medical information for a range of purposes, plus collaborative medical decision-making [63], medical informatics and **diagnosis** (S.J. [61]), and fraud prevention [47]. Because of the blurring of the borders between the health system, wellness sectors, and mobile phones, researchers must recognize such beneficiaries ensuring data is accessed by the relevant authority. Moreover, the important thing is maintaining the integrity of data following ethical and legal bindings. As a result, scholars must concentrate to understand the perspective of the user about the perceived advantages and costs of engaging in a blockchain system. It can assist to identify and remove obstacles in the widespread application and use of blockchain.

Ethical & legal consideration. Blockchain applications are addressing critical issues for example authentication, interoperability, and safe sharing of medical data [49, 118–120]. Regardless of the increased emphasis on the blockchain, the acceptance of such concerns may be regarded as a remarkable barrier to its extensive adoption. More emphasis should be placed on regulatory compliance [100] and ethical recommendation for issues like control of ownership and access of patient data [99]. We propose that future scholars take a multidisciplinary approach to determine avenues for resolving ethical and legal compliance issues in multi-national or cross-institutional contexts for blockchain adoption. We also argue that there is a need to positively impact the public and appease regulatory agencies by deliberating and highlighting the critical benefits derived by using technology based on blockchain.

Directions for future research

According to the SLR, we provide a brief summarization of the thematic problems that would require attention from future researchers:

Deployment of holistic view. In case it is critical to find solutions to security and performance-related problems, like interoperability [67] and access-control [68], we argue that scholars must take a broader view of blockchain adoption. This is critical to creating holistic, legally, and ethically compliant [21, 30], robust data management, and authentication procedures in e-health ecosystems [33]. Furthermore [36], argues that context variables like people and culture may play an important role in the development of new technologies. Eventually, we suggest testing blockchain-based electronic health ecosystems in cross-institutional and cross-national contexts to build tailored context-based healthcare solutions in collaborating with different organizations inside the healthcare space, such as research medical centers [60].

Optimization of the architecture. Scholars might focus on improving the efficiency and performance of proposed designs to account for the higher transaction rates which may be expected if blockchain is integrated into healthcare operations in the future [113]. That can be accomplished by dealing with network congestion [69], scalability [99], throughput [76], and bandwidth issues [22].

Data protection & legal compliance. Addressing data, plus user privacy and legal problems will be an important area of future research [21, 30, 53]. These can be directly tackled by designing blockchain protocols in handling healthcare records that can be enforceable by smart-contract [36] and compliant with data and privacy protection regulations, for example, Health Insurance Portability and Accountability Act [31, 36, 53].

Other technologies integration. For improved functionality, deployment of block-chain might be advantageous by the technology with business processes in healthcare [36]. For example, researchers can concentrate to advance the incorporation of edge computing, AI, and ML through blockchain health service ecosystems in developing an improved anticipatory analytic model to provide customized health treatment

and diagnostics (e.g. [52, 63, 64]). Furthermore, research may aim to improve accessibility, remote control, and emergency services via the integration of sensors based on IoT. Furthermore, we propose two additional potential directions for future scholars to extend the existing scope of academic boundaries in this sector. First of all, it proposes the requirement to understand the implications of blockchain deployment in more niches in healthcare, but related fields i.e. managing the digital rights of users' [13], drug prescription management [11], and prescription fraud prevention [47]. Furthermore, the research could be conducted to investigate the implications of blockchain usage across the whole health system supply and value chain. It can help scholars better understand user-related interoperability problems and additionally enables creating standard protocols to use systems working under the blockchain.

Conclusion

This research study is designed to understand completely the application of blockchain in the domain of healthcare. To achieve this goal, SLRs were conducted on nine highly regarded databases using particular protocols to pick out relevant articles for review. The outcomes were used, to sum up, current knowledge on applications of blockchain in the specific sector of medical care, but to also summarize past and the present academic research theme trends in this field. Future research possibilities have been showcased in the form of a synthesized framework created by combining insights from existing restrictions, suggestions, and emerging gaps in current knowledge observed throughout this review.

Supporting Information

S1 Checklist. PRISMA 2020 checklist.
(DOCX)

Author Contributions

Conceptualization: Hassaan Malik, Umair Bashir, Aiesha Ahmad, Maheen Ilyas, Muhammad Imran Ali Khan.
Data curation: Huma Saeed, Hassaan Malik, Aiesha Ahmad, Shafia Riaz, Wajahat Anwaar Bukhari, Muhammad Imran Ali Khan.
Formal analysis: Umair Bashir, Aiesha Ahmad, Maheen Ilyas, Wajahat Anwaar Bukhari.
Funding acquisition: Shafia Riaz.
Investigation: Aiesha Ahmad.
Methodology: Huma Saeed, Aiesha Ahmad, Shafia Riaz, Maheen Ilyas.
Resources: Muhammad Imran Ali Khan.
Supervision: Hassaan Malik.
Writing – original draft: Huma Saeed, Hassaan Malik, Aiesha Ahmad, Maheen Ilyas.
Writing – review & editing: Umair Bashir, Wajahat Anwaar Bukhari.

References

1. Malik H., Farooq M. S., Khelifi A., Abid A., Nasir Qureshi J. and Hussain M., "A Comparison of Transfer Learning Performance Versus Health Experts in Disease Diagnosis From Medical Imaging," in IEEE Access, vol. 8, pp. 139367–139386, 2020, https://doi.org/10.1109/ACCESS.2020.3004766
2. Campbell S. M., Roland M. O., & Buetow S. A. (2000). Defining quality of care. Social science & medicine, 51(11), 1611–1625. https://doi.org/10.1016/s0277-9536(00)00057-5 PMID:11072882

3. Liao L., Chen M., Rodrigues J. J., Lai X., & Vuong S. (2012). A novel web-enabled healthcare solution on HealthVault system. journal of medical systems, 36(3), 1095–1105. https://doi.org/10.1007/s10916-010-9572-2 PMID:20703631
4. Devadass L., Sekaran S. S., & Thinakaran R. (2017). Cloud computing in healthcare. International Journal of Students' Research in Technology & Management, 5(1), 25–31.
5. Talesh S. A. (2018). Data breach, privacy, and cyber insurance: How insurance companies act as "compliance managers" for businesses. Law & Social Inquiry, 43(2), 417–440.
6. McCoy T. H., & Perlis R. H. (2018). Temporal trends and characteristics of reportable health data breaches, 2010–2017. Jama, 320(12), 1282–1284. https://doi.org/10.1001/jama.2018.9222 PMID:30264106
7. Meinert E., Alturkistani A., Foley K., Osama T., Car J., Majeed F., et al. Implementation of blockchains in healthcare: protocol for a systematic review.
8. Rupasinghe T., Burstein F., Rudolph C., & Strange S. (2019, January). Towards a blockchain based fall prediction model for aged care. In Proceedings of the Australasian Computer Science Week Multiconference (pp. 1–10).
9. Angraal S., Krumholz H. M., & Schulz W. L. (2017). Blockchain Technology: Applications in Health Care. Circulation. Cardiovascular Quality and Outcomes, 10 (9). https://doi.org/10.1161/CIRCOUTCOMES.117.003800 PMID:28912202
10. Hölbl M., Kompara M., Kamišalić A., & Nemec Zlatolas L. (2018). A systematic review of the use of blockchain in healthcare. Symmetry, 10(10), 470.
11. Agbo C. C., Mahmoud Q. H., & Eklund J. M. (2019, June). Blockchain technology in healthcare: a systematic review. In Healthcare (Vol. 7, No. 2, p. 56). Multidisciplinary Digital Publishing Institute.
12. O'donoghue O., Vazirani A. A., Brindley D., & Meinert E. (2019). Design choices and trade-offs in health care blockchain implementations: systematic review. Journal of medical Internet research, 21(5), e12426. https://doi.org/10.2196/12426 PMID:31094344
13. Abou Jaoude J., & Saade R. G. (2019). Blockchain applications–usage in different domains. IEEE Access, 7, 45360–45381.
14. Hasselgren A., Kralevska K., Gligoroski D., Pedersen S. A., & Faxvaag A. (2020). Blockchain in healthcare and health sciences—A scoping review. International Journal of Medical Informatics, 134, 104040. https://doi.org/10.1016/j.ijmedinf.2019.104040 PMID:31865055
15. Risius M., & Spohrer K. (2017). A blockchain research framework. Bus Inf Syst Eng 59: 385–409.
16. Ozdagoglu G., Damar M., & Ozdagoglu A. (2020). The State of the art in blockchain research (2013–2018): scientometrics of the related papers in web of science and scopus. In Digital Business Strategies in Blockchain Ecosystems (pp. 569–599). Springer, Cham.
17. Kitchenham B., Brereton O. P., Budgen D., Turner M., Bailey J., & Linkman S. (2009). Systematic literature reviews in software engineering–a systematic literature review. Information and software technology, 51(1), 7–15.
18. Aznoli F., & Navimipour N. J. (2017). Cloud services recommendation: Reviewing the recent advances and suggesting the future research directions. Journal of Network and Computer Applications, 77, 73–86.

19. Gopalakrishnan S., & Ganeshkumar P. (2013). Systematic reviews and meta-analysis: understanding the best evidence in primary healthcare. Journal of family medicine and primary care, 2(1), 9. https://doi.org/10.4103/2249-4863.109934 PMID:24479036

20. Mendling J., Weber I., Aalst W. V. D., Brocke J. V., Cabanillas C., Daniel F., et al. (2018). Blockchains for business process management-challenges and opportunities. ACM Transactions on Management Information Systems (TMIS), 9(1), 1–16.

21. Kuo T. T., Gabriel R. A., & Ohno-Machado L. (2019). Fair compute loads enabled by blockchain: sharing models by alternating client and server roles. Journal of the American Medical Informatics Association, 26(5), 392–403. https://doi.org/10.1093/jamia/ocy180 PMID:30892656

22. Swan M. (2015). Blockchain: Blueprint for a new economy. "O'Reilly Media, Inc.".

23. Efanov D., & Roschin P. (2018). The all-pervasiveness of the blockchain technology. Procedia computer science, 123, 116–121.

24. Miau S., & Yang J. M. (2018). Bibliometrics-based evaluation of the Blockchain research trend: 2008–March 2017. Technology Analysis & Strategic Management, 30(9), 1029–1045.

25. Yli-Huumo J., Ko D., Choi S., Park S., & Smolander K. (2016). Where is current research on blockchain technology?—a systematic review. PloS one, 11(10), e0163477. https://doi.org/10.1371/journal.pone.0163477 PMID:27695049

26. Angelis J., & da Silva E. R. (2019). Blockchain adoption: A value driver perspective. Business Horizons, 62(3), 307–314.

27. Silva C. A., Aquino G. S., Melo S. R., & Egídio D. J. (2019). A fog computing-based architecture for medical records management. Wireless Communications and Mobile Computing, 2019.

28. Zhang P., White J., Schmidt D. C., Lenz G., & Rosenbloom S. T. (2018). FHIRChain: applying blockchain to securely and scalably share clinical data. Computational and structural biotechnology journal, 16, 267–278. https://doi.org/10.1016/j.csbj.2018.07.004 PMID:30108685

29. Mettler, M. (2016, September). Blockchain technology in healthcare: The revolution starts here. In 2016 IEEE 18th international conference on e-health networking, applications and services (Healthcom) (pp. 1–3). IEEE.

30. Kuo T. T., Kim H. E., & Ohno-Machado L. (2017). Blockchain distributed ledger technologies for biomedical and health care applications. Journal of the American Medical Informatics Association, 24(6), 1211–1220. https://doi.org/10.1093/jamia/ocx068 PMID:29016974

31. Soltanisehat L., Alizadeh R., Hao H., & Choo K. K. R. (2020). Technical, temporal, and spatial research challenges and opportunities in blockchain-based healthcare: A systematic literature review. IEEE Transactions on Engineering Management.

32. Cios, K. J., Krawczyk, B., Cios, J., & Staley, K. J. (2019). Uniqueness of Medical Data Mining: How the new technologies and data they generate are transforming medicine. arXiv preprint arXiv:1905.09203.

33. Ji Y., Zhang J., Ma J., Yang C., & Yao X. (2018). BMPLS: Blockchain-based multi-level privacy-preserving location sharing scheme for telecare medical information systems. Journal of medical systems, 42(8), 1–13. https://doi.org/10.1007/s10916-018-0998-2 PMID:29961160

34. Ito, K., Tago, K., & Jin, Q. (2018, October). i-Blockchain: a blockchain-empowered individual-centric framework for privacy-preserved use of personal health data. In 2018 9th International Conference on Information Technology in Medicine and Education (ITME) (pp. 829–833). IEEE.

35. Iansiti, M., & Lakhani, K. R. (2017). Do Not Copy or Post.

36. Pawczuk, L., Massey, R., & Holdowsky, J. (2019). Deloitte's 2019 Global Blockchain Survey-Blockchain gets down to business. Deloitte Insights, 6.

37. Afrooz S., & Navimipour N. J. (2017). Memory designing using quantum-dot cellular automata: systematic literature review, classification and current trends. Journal of Circuits, Systems and Computers, 26(12), 1730004.

38. Ahmad M. O., Dennehy D., Conboy K., & Oivo M. (2018). Kanban in software engineering: A systematic mapping study. Journal of Systems and Software, 137, 96–113.

39. Mehta N., & Pandit A. (2018). Concurrence of big data analytics and healthcare: A systematic review. International journal of medical informatics, 114, 57–65. https://doi.org/10.1016/j.ijmedinf.2018.03.013 PMID:29673604

40. Zhang R., Simon G., & Yu F. (2017). Advancing Alzheimer's research: A review of big data promises. International journal of medical informatics, 106, 48–56. https://doi.org/10.1016/j.ijmedinf.2017.07.002 PMID:28870383

41. Behera R. K., Bala P. K., & Dhir A. (2019). The emerging role of cognitive computing in healthcare: a systematic literature review. International journal of medical informatics, 129, 154–166. https://doi.org/10.1016/j.ijmedinf.2019.04.024 PMID:31445250

42. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. PLOS Medicine 2021;18(3):e1003583. https://doi.org/10.1371/journal.pmed.1003583 PMID:33780438

43. Yaqoob I., Salah K., Jayaraman R., & Al-Hammadi Y. (2021). Blockchain for healthcare data management: Opportunities, challenges, and future recommendations. Neural Computing and Applications, 1–16.

44. Du X., Chen B., Ma M., & Zhang Y. (2021). Research on the Application of Blockchain in Smart Healthcare: Constructing a Hierarchical Framework. Journal of Healthcare Engineering, 2021. https://doi.org/10.1155/2021/6698122 PMID:33505644

45. Badr S., Gomaa I., & Abd-Elrahman E. (2018). Multi-tier blockchain framework for IoT-EHRs systems. Procedia Computer Science, 141, 159–166.

46. Brogan J., Baskaran I., & Ramachandran N. (2018). Authenticating health activity data using distributed ledger technologies. Computational and structural biotechnology journal, 16, 257–266. https://doi.org/10.1016/j.csbj.2018.06.004 PMID:30101004

47. Casado-Vara R., & Corchado J. (2019). Distributed e-health wide-world accounting ledger via blockchain. Journal of Intelligent & Fuzzy Systems, 36(3), 2381–2386.

48. Dagher G. G., Mohler J., Milojkovic M., & Marella P. B. (2018). Ancile: Privacy-preserving framework for access control and interoperability of electronic health records using blockchain technology. Sustainable cities and society, 39, 283–297.

49. Dhagarra D., Goswami M., Sarma P. R. S., & Choudhury A. (2019). Big Data and blockchain supported conceptual model for enhanced healthcare coverage: The Indian context. Business Process Management Journal.

50. Dwivedi A. D., Srivastava G., Dhar S., & Singh R. (2019). A decentralized privacy-preserving healthcare blockchain for IoT. Sensors, 19(2), 326. https://doi.org/10.3390/s19020326 PMID:30650612

51. Fan K., Wang S., Ren Y., Yang K., Yan Z., Li H., et al. (2018). Blockchain-based secure time protection scheme in IoT. IEEE Internet of Things Journal, 6(3), 4671–4679.

52. Firdaus A., Anuar N. B., Ab Razak M. F., Hashem I. A. T., Bachok S., & Sangaiah A. K. (2018). Root exploit detection and features optimization: mobile device and blockchain based medical data management. Journal of medical systems, 42(6), 1–23. https://doi.org/10.1007/s10916-018-0966-x PMID:29728780

53. Griggs K. N., Ossipova O., Kohlios C. P., Baccarini A. N., Howson E. A., & Hayajneh T. (2018). Healthcare blockchain system using smart contracts for secure automated remote patient monitoring. Journal of medical systems, 42(7), 1–7. https://doi.org/10.1007/s10916-018-0982-x PMID:29876661

54. Guo R., Shi H., Zhao Q., & Zheng D. (2018). Secure attribute-based signature scheme with multiple authorities for blockchain in electronic health records systems. IEEE access, 6, 11676–11686.

55. Hussein A. F., ArunKumar N., Ramirez-Gonzalez G., Abdulhay E., Tavares J. M. R., & de Albuquerque V. H. C. (2018). A medical records managing and securing blockchain based system supported by a genetic algorithm and discrete wavelet transform. Cognitive Systems Research, 52, 1–11.

56. Hyla T., & Pejaś J. (2019). eHealth integrity model based on permissioned blockchain. Future Internet, 11(3), 76.

57. Islam N., Faheem Y., Din I. U., Talha M., Guizani M., & Khalil M. (2019). A blockchain-based fog computing framework for activity recognition as an application to e-Healthcare services. Future Generation Computer Systems, 100, 569–578.

58. Jamil F., Hang L., Kim K., & Kim D. (2019). A novel medical blockchain model for drug supply chain integrity management in a smart hospital. Electronics, 8(5), 505.

59. Kaur H., Alam M. A., Jameel R., Mourya A. K., & Chang V. (2018). A proposed solution and future direction for blockchain-based heterogeneous medicare data in cloud environment. Journal of medical systems, 42(8), 1–11. https://doi.org/10.1007/s10916-018-1007-5 PMID:29987560

60. Lee S. H., & Yang C. S. (2018). Fingernail analysis management system using microscopy sensor and blockchain technology. International Journal of Distributed Sensor Networks, 14(3), 1550147718767044.

61. Lee S. J., Cho G. Y., Ikeno F., & Lee T. R. (2018). BAQALC: blockchain applied lossless efficient transmission of DNA sequencing data for next generation medical informatics. applied sciences, 8(9), 1471.

62. Li H., Zhu L., Shen M., Gao F., Tao X., & Liu S. (2018). Blockchain-based data preservation system for medical data. Journal of medical systems, 42(8), 1–13. https://doi.org/10.1007/s10916-018-0997-3 PMID:29956058

63. Li X., Huang X., Li C., Yu R., & Shu L. (2019). EdgeCare: leveraging edge computing for collaborative data management in mobile healthcare systems. IEEE Access, 7, 22011–22025.

64. Mamoshina P., Ojomoko L., Yanovich Y., Ostrovski A., Botezatu A., Prikhodko P., et al. (2018). Converging blockchain and next-generation artificial intelligence technologies to decentralize and accelerate biomedical research and healthcare. Oncotarget, 9(5), 5665. https://doi.org/10.18632/oncotarget.22345 PMID:29464026

65. Steinfeld B., Scott J., Vilander G., Marx L., Quirk M., Lindberg J., et al. (2015). The role of lean process improvement in implementation of evidence-based practices in behavioral health care. The Journal of Behavioral Health Services & Research, 42(4), 504–518. https://doi.org/10.1007/s11414-013-9386-3 PMID:24464179

66. Nguyen D. C., Pathirana P. N., Ding M., & Seneviratne A. (2019). Blockchain for secure ehrs sharing of mobile cloud based e-health systems. IEEE access, 7, 66792–66806.

67. Al Omar A., Bhuiyan M. Z. A., Basu A., Kiyomoto S., & Rahman M. S. (2019). Privacy-friendly platform for healthcare data in cloud based on blockchain environment. Future generation computer systems, 95, 511–521.

68. Rahmadika S., & Rhee K. H. (2019). Toward privacy-preserving shared storage in untrusted blockchain P2P networks. Wireless Communications and Mobile Computing, 2019.

69. Shen B., Guo J., & Yang Y. (2019). MedChain: Efficient healthcare data sharing via blockchain. Applied sciences, 9(6), 1207.

70. Tian H., He J., & Ding Y. (2019). Medical data management on blockchain with privacy. Journal of medical systems, 43(2), 1–6. https://doi.org/10.1007/s10916-018-1144-x PMID:30603816

71. Uddin M. A., Stranieri A., Gondal I., & Balasubramanian V. (2018). Continuous patient monitoring with a patient centric agent: A block architecture. IEEE Access, 6, 32700–32726.

72. Crosby M., Pattanayak P., Verma S., & Kalyanaraman V. (2016). Blockchain technology: Beyond bitcoin. Applied Innovation, 2(6–10), 71.

73. Wong D. R., Bhattacharya S., & Butte A. J. (2019). Prototype of running clinical trials in an untrustworthy environment using blockchain. Nature communications, 10(1), 1–8.

74. Yang J., Onik M. M. H., Lee N. Y., Ahmed M., & Kim C. S. (2019). Proof-of-familiarity: a privacy-preserved blockchain scheme for collaborative medical decision-making. Applied Sciences, 9(7), 1370.

75. Zhang A., & Lin X. (2018). Towards secure and privacy-preserving data sharing in e-health systems via consortium blockchain. Journal of medical systems, 42(8), 1–18. https://doi.org/10.1007/s10916-018-0995-5 PMID:29956061

76. Zheng X., Sun S., Mukkamala R. R., Vatrapu R., & Ordieres-Meré J. (2019). Accelerating health data sharing: A solution based on the internet of things and distributed ledger technologies. Journal of medical Internet research, 21(6), e13583. https://doi.org/10.2196/13583 PMID:31172963

77. Zhou L., Wang L., & Sun Y. (2018). MIStore: a blockchain-based medical insurance storage system. Journal of medical systems, 42(8), 1–17. https://doi.org/10.1007/s10916-018-0996-4 PMID:29968202

78. Tandon A., Dhir A., Islam N., & Mäntymäki M. (2020). Blockchain in healthcare: A systematic literature review, synthesizing framework and future research agenda. Computers in Industry, 122, 103290.

79. Chukwu E., & Garg L. (2020). A systematic review of blockchain in healthcare: frameworks, prototypes, and implementations. IEEE Access, 8, 21196–21214.

80. Celesti A., Ruggeri A., Fazio M., Galletta A., Villari M., & Romano A. (2020). Blockchain-based healthcare workflow for tele-medical laboratory in federated hospital IoT clouds. Sensors, 20(9), 2590. https://doi.org/10.3390/s20092590 PMID:32370129

81. Fusco A., Dicuonzo G., Dell'Atti V., & Tatullo M. (2020). Blockchain in healthcare: Insights on COVID-19. International Journal of Environmental Research and Public Health, 17(19), 7167. https://doi.org/10.3390/ijerph17197167 PMID:33007951

82. Agbo C. C., & Mahmoud Q. H. (2020). Blockchain in Healthcare: Opportunities, Challenges, and Possible Solutions. International Journal of Healthcare Information Systems and Informatics (IJHISI), 15(3), 82–97.

83. Mazlan A. A., Daud S. M., Sam S. M., Abas H., Rasid S. Z. A., & Yusof M. F. (2020). Scalability challenges in healthcare blockchain system—a systematic review. IEEE Access, 8, 23663–23673.

84. Khatoon A. (2020). A blockchain-based smart contract system for healthcare management. Electronics, 9(1), 94.

85. Dubovitskaya, A., Xu, Z., Ryu, S., Schumacher, M., & Wang, F. (2017). Secure and trustable electronic medical records sharing using blockchain. In AMIA annual symposium proceedings (Vol. 2017, p. 650). American Medical Informatics Association.

86. Azaria, A., Ekblaw, A., Vieira, T., & Lippman, A. (2016, August). Medrec: Using blockchain for medical data access and permission management. In 2016 2nd international conference on open and big data (OBD) (pp. 25–30). IEEE.

87. Alwen, J., Blocki, J., & Harsha, B. (2017, October). Practical graphs for optimal side-channel resistant memory-hard functions. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 1001–1017).

88. Lei A., Cruickshank H., Cao Y., Asuquo P., Ogah C. P. A., & Sun Z. (2017). Blockchain-based dynamic key management for heterogeneous intelligent transportation systems. IEEE Internet of Things Journal, 4(6), 1832–1843.

89. Huang J., He D., Obaidat M. S., Vijayakumar P., Luo M., & Choo K. K. R. (2021). The Application of the Blockchain Technology in Voting Systems: A Review. ACM Computing Surveys (CSUR), 54(3), 1–28.

90. Zhou T., Shen J., He D., Vijayakumar P., & Kumar N. (2020). Human-in-the-loop-aided privacy-preserving scheme for smart healthcare. *IEEE Transactions on Emerging Topics in Computational Intelligence*.

91. Qian Y., Shen J., Vijayakumar P., & Sharma P. K. (2021). Profile Matching for IoMT: A Verifiable Private Set Intersection Scheme. IEEE Journal of Biomedical and Health Informatics. https://doi.org/10.1109/JBHI.2021.3088289 PMID:34111016

92. Lin C., He D., Kumar N., Huang X., Vijayakumar P., & Choo K. K. R. (2019). Homechain: A blockchain-based secure mutual authentication system for smart homes. *IEEE Internet of Things Journal*, 7(2), 818–829.

93. Sandgaard, J., & Wishstar, S. (2018). MedChain. September, 30, 2018.

94. Albeyatti, A. (2018). Meddicalchain. September, 30, 2018.

95. Ganesan R., Devi T., Kumar S. R., & Gayathri N. (2022). Securing Healthcare Information Using Blockchain Technology: A Deep Insight. In Blockchain Security in Cloud Computing (pp. 253–263). Springer, Cham.

96. Xu M., Chen X. & Kou G. A systematic review of blockchain. *Financ Innov* 5, 27 (2019).

97. Hoy M. B. (2017). An introduction to the blockchain and its implications for libraries and medicine. Medical reference services quarterly, 36(3), 273–279. https://doi.org/10.1080/02763869.2017.1332261 PMID:28714815

98. Gilboy M. B., Heinerichs S., & Pazzaglia G. (2015). Enhancing student engagement using the flipped classroom. Journal of nutrition education and behavior, 47(1), 109–114. https://doi.org/10.1016/j.jneb.2014.08.008 PMID:25262529

99. Zheng Z., Xie S., Dai H. N., Chen X., & Wang H. (2018). Blockchain challenges and opportunities: A survey. International Journal of Web and Grid Services, 14(4), 352–375.

100. Gutiérrez O., Romero G., Pérez L., Salazar A., Charris M., & Wightman P. (2020). HealthyBlock: Blockchain-Based IT Architecture for Electronic Medical Records Resilient to Connectivity Failures. *International Journal of Environmental Research and Public Health*, 17(19), 7132. https://doi.org/10.3390/ijerph17197132 PMID:33003452

101. Roehrs A., Da Costa C. A., & da Rosa Righi R. (2017). OmniPHR: A distributed architecture model to integrate personal health records. Journal of biomedical informatics, 71, 70–81. https://doi.org/10.1016/j.jbi.2017.05.012 PMID:28545835

102. Tosh, D. K., Shetty, S., Liang, X., Kamhoua, C. A., Kwiat, K. A., & Njilla, L. (2017, May). Security implications of blockchain cloud with analysis of block withholding attack. In 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID) (pp. 458–467). IEEE.

103. Wood G. (2014). Ethereum: A secure decentralised generalised transaction ledger. Ethereum project yellow paper, 151(2014), 1–32.

104. Wahsheh, H. A., & Al-Zahrani, M. S. (2021, May). Secure and Usable QR Codes for Healthcare Systems: The Case of Covid-19 Pandemic. In 2021 12th International Conference on Information and Communication Systems (ICICS) (pp. 324–329). IEEE.

105. Tikhomirov, S. (2017, October). Ethereum: state of knowledge and research perspectives. In International Symposium on Foundations and Practice of Security (pp. 206–221). Springer, Cham.

106. O'Dwyer, K. J., & Malone, D. (2014). Bitcoin mining and its energy footprint.

107. McFarlane, C., Beer, M., Brown, J., & Prendergast, N. (2017). Patientory: A Healthcare Peer-to-Peer EMR Storage Network v1. 1, Whitepaper. May.

108. Benchoufi M., Porcher R., & Ravaud P. (2017). Blockchain protocols in clinical trials: Transparency and traceability of consent. F1000Research, 6.

109. Powles J., & Hodson H. (2017). Google DeepMind and healthcare in an age of algorithms. Health and technology, 7(4), 351–367. https://doi.org/10.1007/s12553-017-0179-1 PMID:29308344

110. Wachter R. (2016). Making IT work: harnessing the power of health information technology to improve care in England. London, UK: Department of Health.

111. Horblyuk R., Kaneta K., McMillen G. L., Mullins C., OBrien T. M., & Roy A. (2012). Out of control: little-used clinical assets are draining healthcare budgets: little-used mobile clinical equipment is a significant expense for healthcare organizations, but cutting costs involves more than inventory reduction. Healthcare Financial Management, 66(7), 64–69.

112. Yue X., Wang H., Jin D., Li M., & Jiang W. (2016). Healthcare data gateways: found healthcare intelligence on blockchain with novel privacy risk control. Journal of medical systems, 40(10), 1–8. https://doi.org/10.1007/s10916-016-0574-6 PMID:27565509

113. Li Z., Wang W. M., Liu G., Liu L., He J., & Huang G. Q. (2018). Toward open manufacturing: A cross-enterprises knowledge and services exchange framework based on blockchain and edge computing. Industrial Management & Data Systems.

114. Xia Q. I., Sifah E. B., Asamoah K. O., Gao J., Du X., & Guizani M. (2017). MeDShare: Trust-less medical data sharing among cloud service providers via blockchain. IEEE Access, 5, 14757–14767.

115. Noh S. W., Sur C., Park Y. H., Shin S. U., & Rhee K. H. (2017). Blockchain-based user-centric records management system. Int J Control Autom, 10(11), 133–144.

116. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019; 366: l4898. https://doi.org/10.1136/bmj.l4898 PMID:31462531

117. Higgins JPT, Sterne JAC, Savović J, Page MJ, Hróbjartsson A, Boutron I, et al. (editors). Cochrane Methods. *Cochrane Database of Systematic Reviews* 2016, Issue 10 (Suppl 1).

118. Bocek, T., Rodrigues, B. B., Strasser, T., & Stiller, B. (2017, May). Blockchains everywhere-a use-case of blockchains in the pharma supply-chain. In 2017 IFIP/IEEE symposium on integrated network and service management (IM) (pp. 772–777). IEEE.

119. Dimitrov D. V. (2019). Blockchain applications for healthcare data management. Healthcare informatics research, 25(1), 51–56. PMID:30788182

120. Casino F., Dasaklis T. K., & Patsakis C. (2019). A systematic literature review of blockchain-based applications: Current status, classification and open issues. Telematics and informatics, 36, 55–81.

7.8 Case Study 3

RESEARCH ARTICLE

Assessing the impact of technological change on similar occupations: Implications for employment alternatives

Karine Torosyan[1,2], Sicheng Wang [3]*, Elizabeth A. Mack[1], Jenna A. Van Fossen[4], Nathan Baker[5]

1 Department of Geography, Environment and Spatial Sciences, Michigan State University, East Lansing, Michigan, United States of America, 2 International School of Economics in Tbilisi (ISET), Tbilisi State University, Tbilisi, Georgia, 3 Department of Geography, University of South Carolina, Columbia, South Carolina, United States of America, 4 Department of Psychology, Clemson University, Clemson, South Carolina, United States of America, 5 Department of Psychology, Michigan State University, East Lansing, Michigan, United States of America

*sichengw@mailbox.sc.edu

Abstract

Background

The fast-changing labor market highlights the need for an in-depth understanding of occupational mobility impacted by technological change. However, we lack a multidimensional classification scheme that considers similarities of occupations comprehensively, which prevents us from predicting employment trends and mobility across occupations. This study fills the gap by examining employment trends based on similarities between occupations.

Method

We first demonstrated a new method that clusters 756 occupation titles based on knowledge, skills, abilities, education, experience, training, activities, values, and interests. We used the Principal Component Analysis to categorize occupations in the Standard Occupational Classification, which is grouped into a four-level hierarchy. Then, we paired the occupation clusters with the occupational employment projections provided by the U.S. Bureau of Labor Statistics. We analyzed how employment would change and what factors affect the employment changes within occupation groups. Particularly, we specified factors related to technological changes.

Results

The results reveal that technological change accounts for significant job losses in some clusters. This poses occupational mobility challenges for workers in these jobs at present. Job losses for nearly 60% of current employment will occur in low-skill, low-wage occupational groups. Meanwhile, many mid-skilled and highly skilled jobs are projected to grow in the next ten years.

Conclusion

Our results demonstrate the utility of our occupational classification scheme. Furthermore, it suggests a critical need for skills upgrading and workforce development for workers in declining jobs. Special attention should be paid to vulnerable workers, such as older individuals and minorities.

https://doi.org/10.1371/journal.pone.0291428 Used with CC 4.0 license

Introduction

Technological change is a disruptive force on labor markets, and studies find that jobs are becoming increasingly polarized into low and high-skilled jobs [1–4]. Recent technological innovations related to robotics and artificial intelligence are anticipated to have particularly disruptive impacts on the workforce [5]. By 2025, projections indicate that these changes will create 97 million new jobs and destroy 85 million current jobs globally [6]. A concern about this recent wave of new technologies is that they will begin to replace high-skill, high-wage jobs [7, 8]; thereby placing previously stable work in a precarious position.

Two economic theories help explain this workforce polarization. The theory of skills-biased technological change (SBTC) predicts that the demand for highly skilled workers, and the wages they earn, rises during periods of technological change [1]. Based on this theory, technology complements highly skilled jobs and substitutes for jobs with low and medium skills, leading to a rise in income inequality [3]. Instead, the theory of routine-biased technological change (RBTC) [9] offers a more nuanced perspective on how advances in computing technologies replace human labor in routine tasks, which can be decomposed into step-by-step procedures or rules. Thus, the RBTC framework helps explain the disappearance of the "middle" in both manual and cognitive-related jobs that are replaceable by machines due to the predictable, routine aspects of work [2]. On the other hand, non-routine jobs, either highly-skilled or unskilled, are likely to increase as complements to technologies [2]. Therefore, the projection of job demand requires a comprehensive assessment of all occupational attributes beyond skills, including knowledge, experience, training, activities, interests, and values.

One of the important aspects of technological change in the labor market to consider is its impact on occupational mobility or the ability of workers to move between jobs. This is important to evaluate for workers who are displaced by technology. Based on the theory of RBTC, one would expect workers employed in non-routine jobs with a high cognitive task content to be more mobile than workers in both routine manual and routine cognitive work. Cortes [10] evaluated occupational mobility in the US between 1976 and 2007 and categorized occupations into the three groups proposed by Autor et al. [9]. They find routine workers experienced a fall in their wage premiums while people who moved out of routine jobs were more likely to experience wage growth. Maczulskij [11] examined the occupational mobility of mid-skilled routine workers in Finland and found differences between routine cognitive workers and routine manual workers. In particular, they find that routine cognitive workers can move to abstract tasks while routine manual workers tend to move to service tasks. Goos et al. [12] examine the reemployment prospects of workers after plant closure in Belgium using an RBTC framework. They find evidence in support of this theory. The reemployment prospects for workers performing non-routine tasks and with digital skills were better than for workers accustomed to performing routine tasks. That said, few studies have examined the impact of technological change on occupational mobility, particularly in a U.S. context, and more work needs to be done in this area [11, 13].

Analyses of occupational mobility classify occupations into groups based on their similarity in characteristics (e.g., wages, skills, tasks) [12, 14–24]. The logic behind the creation of these groups is that movement between similar jobs *within* groups is easier than movement *between* groups where jobs are more dissimilar. Thus, the availability of more alternatives within an occupational cluster enhances the number of similar employment alternatives for workers. A common approach of occupational mobility studies is to cluster occupations with a focus on one dimension. For example, some studies categorize occupations based on skills [25], tasks [26], or work values and interests [27–29]. In particular, principal components analysis (PCA) has been used to reduce very large numbers of skills or tasks to assist in classifying jobs [17, 25, 30]. Considering different occupational characteristics, although there has been some dispute over the distinction between tasks and skills, Yamaguchi [31] delineated between tasks and skills and proposed that more observable tasks are indicative of particular skills. It should be noted that there are some existing occupational classification

https://doi.org/10.1371/journal.pone.0291428

systems, such as the International Standard Classification of Occupations (ISCO) and Standard Occupational Classification (SOC). The ISCO system is not recommended for US-based analysis because "it is not flexible enough for US needs" [32]. The SOC system, established by the Office of Management and Budget, is for use in the Federal statistical system in the US [33]. However, the principle of the SOC classifications is based on "work performed" and "the highest level of skills" required in each occupation [32, 33], which does not thoroughly reflect similarities regarding occupational attributes, human capital requirements, and activities. Therefore, SOC classifications may not be sufficient for occupational mobility studies.

The Occupational Classification Network (O*NET) data, which indexes detailed occupational attributes for US occupations, has been widely used in grouping occupations and examining the job characteristics' relatedness [33–43]. Hundreds of attributes, termed as descriptors, are divided into different dimensions, such as knowledge, skills, abilities, education, experience, training, interests, work values, work styles, tasks, technology skills & tools, work activities, and work contents. Each descriptor is rated with a scale. Therefore, analysts can examine the relationships between descriptors based on the values. The data is provided to "help people find the training and jobs they need, and employers the skilled workers" [34]. However, it is difficult to identify alternatives for a worker in a specific occupation because O*NET does not group occupations based on the values of its descriptors. Some studies, such as Feser [35], Koo [36], Nolan et al. [37], Slaper [38] and Burrus et al. [39], used PCA clustering with the Ward algorithm to determine the similarity based on some occupational descriptors in O*NET. Feser [35], Koo [36], and Nolan et al. [37] all created clusters based on the "knowledge" descriptors. Slaper [38] used "knowledge," "skills," and "training" in clustering. Burrus et al. [39] used "knowledge," "skill," "ability," and "work style." Manzella et al. [40] used structured and unstructured factor analysis to group occupations based on "task" descriptors. Yu et al. [41] used PCA with a hierarchical cluster analysis to group occupations at two levels. Their analysis was based on descriptors of "knowledge," "skills," "abilities," and "training," but the higher-level clusters considered only job training requirements. Waters and Shutters [42] and their earlier work [43] created a skills network based on values of six categories of descriptors in O*NET (knowledge, skills, interests, work activities, work styles, and work values) and employment in an occupation in a metropolitan area. Alabdulkareem et al. [44] and Farinha et al. [45] used revealed comparative advantage (RCA, also called a location quotient) to determine whether an O*NET attribute is present or absent in an occupation. RCA helped calculated network-based similarity by examining the relatedness and complementarity of occupational attributes in O*NET. The former study [44] used the importance scale for 161 descriptors of "knowledge," "skills," and "abilities," and the latter [45] used descriptors of "Intermediate Work Activities" (IWA), a sub-category of work activities.

This study develops a nested occupational classification scheme based on the theory of human capital. In Becker's [46] theory of human capital, human capital is defined with multiple dimensions, namely knowledge, skills, abilities, health, and values, with education and training identified as the most important ways to invest in human capital. And indeed, there is ample empirical evidence demonstrating the importance of these factors in determining labor market outcomes, including job placement and mobility [47].

Conceptualizing similarity in terms of skills, tasks, abilities, training, and related information may indicate whether a worker can enter a job. However, it is also valuable to incorporate information on values and interests to get a sense of worker performance and satisfaction in a new job. This is particularly the case given the more recent focus of research on meaningful aspects of work, as people are beginning to view jobs as more than a means of making an income [48, 49]. Interests and values predict long-term performance and turnover in a given job [50–52]. A recent meta-analysis conducted a quantitative review of interests and career choice over 100 years, suggesting that measured interests attain a hit rate of 51% for predicting career choice [53]. The aforementioned studies, such as Shutters and Waters [43] and Waters and Shutters [42] have already included interest elements in occupational clustering. Values, meanwhile, reflect beliefs that motivate evaluations of what environmental characteristics and actions are most preferable [54, 55], influencing how well an employee fits within a given role and, thus, their likelihood of turnover [56]. Determinations about the

https://doi.org/10.1371/journal.pone.0291428

work values that a given occupation satisfies are based on the theory of work adjustment [57], which views values (e.g., achievement, independence) as factors that influence what elements of a given job a worker finds desirable. Similarly, vocational interests reflect preferences for certain types of environments and activities [58]. Holland's [50] RIASEC is the most popular model of vocational interests, proposing that people differ along six stable work interests: realistic, investigative, artistic, social, enterprising, and conventional. Holland postulates that people select careers that fit their personality orientations, as closer alignment between personality and work environment should lead to better job experiences. Unlike knowledge and skills which a worker must acquire when transitioning from one occupation to another, integrating values and interests into occupational clustering can provide job seekers with a sense that the jobs may be personally satisfying, which could enhance their satisfaction, performance, and long-term retention in alternative occupations.

In conclusion, the current literature lacks studies that project employment trends and mobility across occupations based on similarities in multiple factors. Moreover, although prior studies have explored occupation clustering in various approaches, we lack a comprehensive classification scheme for predicting employment trends and occupational mobility driven by technological change. Our paper builds on prior work focused on occupational clustering by creating a new multidimensional classification scheme of jobs. We move beyond prior work focused on occupational clustering to consider the impact of technological change on the size of occupational clusters. This a priori step is important because technological change may eliminate several jobs within occupational clusters, limiting the number of occupational alternatives for workers and their ability to find jobs. Based on human capital theory, we cluster over 700 occupational titles across their knowledge, skills and abilities (KSA), human capital requirements (education, experience, and training (EET)) and work activities, values and interests (AVI). The classification method considers numerous dimensions of potential similarity across occupations, including skills, tasks, human capital-related occupational characteristics (e.g., experience and educational attainment), work values, and interests. Some dimensions, particularly work interests, are important to incorporate in occupational classifications because they have meaningful implications for occupational transitions [53, 59, 60]. The four-level nested hierarchical occupation classifications provide for a finer level of classification. More importantly, we pair this output with information about the drivers of growth or decline for particular occupations from the Bureau of Labor Statistics (BLS) to answer two questions. One, which occupations will grow or decline because of technological change? Two, to what extent does technological change eliminate occupations within clusters? This final question is important for assessing future employment alternatives for workers displaced by technological change. If entire clusters are displaced by technological change, the occupational mobility of displaced workers could be severely limited.

By examining the employment projections for occupations clustered in our analysis, we reveal that job losses of nearly 60% of current employment will occur in low-skill, low-wage occupational groups. On the other hand, primary job gains will be made in high-wage high-educational attainment jobs, which require a Master's degree or higher. Our analysis also high-lights future growth in high-wage, high-educational attainment occupations as well as occupations in the "middle" portion of the spectrum. These "middle" occupations include various types of technician, programming, detective and investigative work, financial examiners and police work. While work in these occupations could be somewhat routine, they also require a human reasoning element that cannot be replaced by machines right now. Much of this job growth is related to Internet-related technologies and the Internet of Things (IoT). This growth in mid-wage jobs suggests that routine-biased technological change (RBTC) is not projected to occur in the short run (i.e., 2029). Moving forward, the wages of these jobs should be tracked to understand long-term trends in wages.

From a policy perspective, the disappearance of work at the lower end of the spectrum means finding a way of retraining workers so they are better equipped to cope with labor market changes wrought by technological change. For a long time, employers in the US have not done an effective job of providing on-the-job training

https://doi.org/10.1371/journal.pone.0291428

[61, 62]. The nested occupational clustering based on O*NET data provides a foundation for developing worker training and certificate programs by identifying knowledge, skills, and educational gaps between declining and growing occupations. Moreover, the growing inequality in educational opportunities throughout childhood and young adulthood has been exacerbated by technological advances [63, 64]. Our analyses of wage patterns and employment trends within occupational clusters shed light on policymaking that aims to mitigate technological gaps in educational opportunities. Finally, integrating information about values and interests satisfied in occupations into groupings of similarity can also provide job seekers with a sense of the occupations that may be personally satisfying, which could increase their satisfaction and long-term retention in alternative occupations.

Materials and methods

Data

To understand the impact of technological change on jobs within occupation clusters, we use information from three data sources. O*NET database (version 25.0) provides hundreds of detailed attributes of occupational characteristics, requirements, and activities for US occupations, described as job descriptors [65]. For every indexed occupation, the descriptors are categorized into a variety of categories, including knowledge, skills, abilities (KSA), education, experience, training (EET), work activities, values, and interests (AVI), as well as tasks, and technology skills and tools [66]. Every category is indexed in a separate data file. For example, the "knowledge" file includes descriptors such as "Administration and Management," "Communications and Media," and "Chemistry." For each occupation, O*NET associates each descriptor with a scale with a numeric range, such as importance, level, relevance, frequency, occupational interest, extent of the activity, and content. Appendices A-C in S2 Appendix list the KSA, EET, and AVI elements from the O*NET database used in our analysis for 220 descriptors. O*NET collects data for over 900 occupations categorized based on the Standard Occupational Classification System (SOC). We work with occupations based on 2018 SOC occupational codes (722 titles). In cases where data for 2018 SOC occupations were not collected (149 titles), we aggregate data from more detailed classification codes (2018 SOC occupational descriptors are computed as the averages of descriptors for lower-level titles included in a given occupation). This produces a final list of 756 occupations to cluster.

After clustering the occupations, we examine occupational wage variation within clusters. We use May 2019 wage and employment estimates from the Occupational Employment Statistics (OES) program from the US Bureau of Labor Statistics (BLS) to examine wages within clusters. The OES survey reports estimated wages by occupation (following the 2010 SOC system), which includes straight-time gross pay and excludes any bonuses and nonwage benefits. An occupational mean annual wage estimate is calculated by summing the annual wages of all the employees in a given occupation and dividing the total wages by the number of employees in that occupation.

We also use data from the Employment Projections (EP) program of the BLS, which publishes 10-year projections of national employment by industry and occupation [67]. The projection is based on analyses of historical and present economic statistics for the labor force, aggregate economic growth, commodity final demand, input-output, industry output and employment, job openings, wage and salary, and factors affecting demand for occupations, including technological innovation, changes in business or production models, product or service replacement, organizational restructuring of work, changes to the size of business establishments, and offshore outsourcing [68]. The EP data used in this study provide projections for 2019–2029. In addition to the projected number of jobs by occupation, EP tables [69] contain a summary of the main factors behind projected changes in reported occupations. This enables us to understand why jobs are projected to grow or decline. A helpful feature of the explanations associated with the job growth or

https://doi.org/10.1371/journal.pone.0291428

decline for each occupation is that the description outlines if the changes are due to changes in technology. We integrate this information with our clustering scheme to analyze technology-driven job dynamics within occupational groups.

Methods

We use the O*NET data release version 25.0 (August 2020) [66] to extract the 220 occupational descriptors discussed in the data section. Many of these descriptors are highly correlated with each other: the pairwise correlation values range from -0.76 to 0.97. To reduce the dimensionality of the information in this database and to produce a set of orthogonal descriptors of occupations, we perform Principal Component Analysis (PCA) using the standardized versions of the original variables and retain the first 72 principal components. These 72 principal components account for more than 90% of the total variation in the data and are used to perform cluster analysis.

To cluster occupations, we use the Ward agglomerative hierarchical clustering technique to create a hierarchy of occupations. This choice of method is driven by the following considerations: a) we aim to create a hierarchy of occupations–from more aggregate to less aggregate, and the use of a hierarchical method helps to achieve this goal (partitioning methods do not produce a clear hierarchy of clusters), b) we want to avoid creating clusters of heterogeneous size–a situation which is typical when using some hierarchical clustering methods, most notably single-and average-linkage methods, but less common with Ward's method. The similarity between observations is determined based on a squared Euclidean distance-based measure.

This selection of the distance measure is based on prior work, which warns against using other measures of distance with Ward's clustering [70–73]. To determine the stopping points for different classification levels, we use the Duda and Hart [74] statistic, which pairs well with agglomerative clustering algorithms [75]. Large values of the Duda-Hart (DH) statistic indicate a distinct cluster structure for the given number of clusters, while small values indicate a less clearly defined cluster structure. Fig 20 graphs the DH statistic based on Ward's hierarchical clustering of occupations. Points of local maxima indicate the most meaningful splits in the occupational hierarchy.

Based on the DH statistics, we select 2 points of local maxima as our stopping points: these are 8-and 17-group splits with the value of DH statistic of 0.89 and 0.92, respectively. Based on these considerations, we construct the following taxonomy of occupations:

- Level 1: 8 Occupational Clusters
- Level 2: 17 Occupational Groups

In the technical appendix (S1 Appendix), we extend the taxonomy of occupations to include a 40-group split (Level 3) and a 59-group split (Level 4), and the complete list of the taxonomy is presented in Appendix D in S2 Appendix.

Our clustering algorithm does not use wages to group occupations, so this is an external variable that we use to assess the quality of our clustering results. The rationale for doing so is that occupations with similar human capital requirements should pay similar wages (e.g., [76]). To assess the homogeneity in wages within clusters, we compute the sum of squares within groups (SSW):

$$SSW_C = \sum_{j=1}^{N_C} \sum_{i=1}^{n_j} \left(w_{ij} - \overline{w}_j \right)^2$$

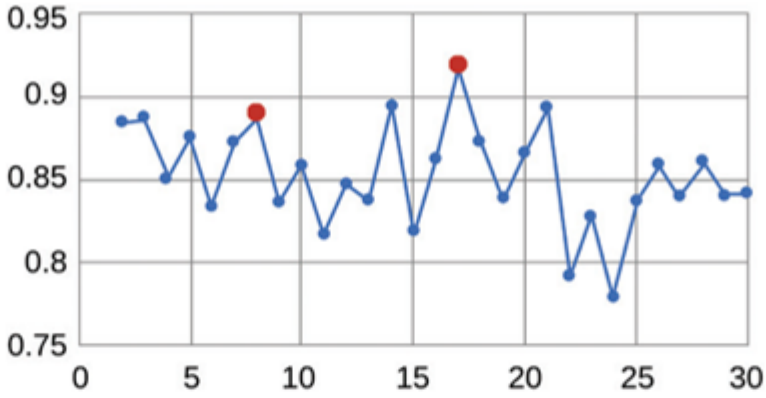


Fig 20. Duda-hart index and points of local maxima for 8 and 17-group splits (in red). Sources: Ward hierarchical clustering method based on 72 principal components derived from 220 occupational descriptors from O*NET.

https://doi.org/10.1371/journal.pone.0291428.g001

where w_{ij} represents the annual mean wage for occupation i within Cluster j, $\bar{w}_j$ is the average wage within Cluster j, nj is the size of Cluster j, and $N_c = 8$ and 17 is the number of clusters at a given classification level c = 1, 2. Next, we estimate the upper bound for the sum of squares within groups (SSW_U) based on a random assignment of wages into the existing cluster structure. This scenario corresponds to the case when the clustering scheme does not explain any variation in wages. The lower bound for the sum of squares within groups (SSW_L) is calculated by partitioning sorted wages into the existing cluster structure. This scenario corresponds to the maximum "fit" given the cluster structure and observed wage data. The following goodness-of-fit measure should reflect where the actual SSW falls between these two extremes:

$$Fit_c = \frac{SSW_{c,U} - SSW_c}{SSW_{c,U} - SSW_{c,L}}$$

Table 6 provides the resulting goodness-of-fit measure for our two levels of classification based on occupation-level wage data from the OES. At the most aggregate level of classification, 38% of the variation in wages is captured by our cluster structure. The fit improves as we move to a finer classification level, reaching 49% for the 17-group solution. Considering wages are not directly utilized in clustering, this level of fit based on wages suggests that our cluster solutions lead to compact groupings of occupations, especially at a finer level of classifications.

To evaluate employment patterns and trends within clusters, we use EP data from the Bureau of Labor Statistics. The EP data estimate the employment trends by occupation in the next decade [67]. The data used in this study were projected employment trends for 2019–2029. There were 162.8 million people employed in 2019 (in occupations covered in BLS projections), and the projected number employed in 2029 is 168.9 million. This is equivalent to a 3.7% growth rate over the entire decade, a very conservative projection. For comparison, the projected growth in jobs for the 2009–2019 period was as high as 13.7% (based on the BLS archive), and the actual growth for the same period was around 14.9% [77]. So, the overall expansion in the employment base in the next decade is projected to be modest. Taking the 3.7% growth rate as a baseline scenario, we identify slower and faster-growing occupations and calculate the negative or positive gaps in the number of jobs within each occupation compared to this baseline scenario. We then aggregate these gaps by occupational groupings to compute the total number of jobs that will be lost and gained within each grouping in the coming decade, compared to the baseline.

For most occupations with sizable changes in employment, the BLS reports the main reasons behind their projected changes [69]. In total, this additional information is available for approximately one-third of all occupations, but together they account for 70% of all projected changes in employment. This information is provided in the form of a textual explanation for each given occupation, and it contains details on the direction of change (negative or positive) and a description of the reasons behind this change. We categorize the range of reasons appearing in the BLS explanations as "related to technology" and "not related to technology." Table 7 lists the types of reasons reported by the BLS, grouped by their relationship to technology. Column 2 of Table 7 provides some descriptions of the reasons driving employment projections. Column 3 shows examples of detailed reasons that fall into these categories. For some occupations, technologies create new tasks and responsibilities, such as handling increasing threats of cyber-attacks, analyzing digital data, or inventing new software and hardware. In other occupations, technologies change the nature of work. For example, some jobs are replaced by automation, AI, or robotics; some jobs based on traditional products or materials will be in lower demand as new products emerge (locksmiths are declining due to the adoption of smart locks); some jobs are in higher demand because of the deployment of new infrastructure (e.g., small commercial satellites and drones); some jobs are displaced because of internet-based platforms (e.g., ride-hailing apps, delivery apps).

Table 6. Wage sums of squares within clusters and goodness of fit.

Level of clustering	SSW			Goodness of Fit
	Actual	Lower bound	Upper bound	
c = 1 (8 clusters)	552,280	13,256	883,236	0.38
c = 2 (17 groups)	448,671	12,946	875,489	0.49

https://doi.org/10.1371/journal.pone.0291428.t001

Table 7. Reasons behind bureau of labor statistics employment projections for change in jobs by occupations.

Relationship to technology	Reason Category (coded from the detailed reasons)	Description of Reason from Bureau of Labor Statistics Projections
Related to technology	Need to manage risks, increase cyber security	Information Security Analyst (15–1212): Share increases as firms increase their IT security staff and capabilities as a response to increasing threats of cyber-attacks aimed at sensitive or financial data.
	Need to handle and/or analyze more/new data	Operations Research Analyst (15–201): share increases as larger amounts of digital and electronic data are collected over the next decade
	Need to develop new software/apps and/or hardware	Computer and Information Research Scientist (15–1221): share increases as demand for digital devices increases
	Need to develop and maintain tools to adjust to new technologies	Industrial Machinery Mechanics (49–9041): share increases as manufacturing automation increases requiring more maintenance and repair
	Job changes because of new infrastructures	Aerospace engineers (17–2011): share increases as aerospace engineers are needed in developing new technologies stemming from the wider commercial use of small satellites and drones.

Relationship to technology	Reason Category (coded from the detailed reasons)	Description of Reason from Bureau of Labor Statistics Projections
	Job changes because of automation, AI, or robotics	Insurance claims and policy processing clerks (43–9041): share decreases as firms increase the use of computer software, enhanced with AI technology, to automate claims and policy processing.
	Job changes because of internet-based apps or platforms	Driver/sales workers (53–3031): share decreases as restaurant food delivery will increasingly be handled by delivery companies that utilize smartphone apps and classify workers as independent contractors.
	Job changes because of innovation or the emergence of products	Locksmiths and safe repairers (49–9094): share decreases as new smart locks, apps on smartphones, key kiosks, and voice–activated locks reduce the need to hire a locksmith to open locks or create key copies.
NOT related to technology	Organizational, industrial, economic restructuring	Cooks, Institution and Cafeteria (35–2012): share decreases as the use of food contractors continues to grow
	Specialization/higher demand for specialists (not due to tech)	Tree Trimmers and Pruners (37–3013): share increases as local governments desire more trees in urban areas
	Bad economic trends and decline in demand for specialists (not due to tech)	Carpet installers (47–2041): share decreases as other floor coverings replace carpet
	Offshoring, outsourcing, occupational substitution	Pharmacy Aides (31–9095): share decreases as pharmacy technicians, cashiers, and clerks take over duties traditionally done by aide
	Demographic change (including aging) and healthier lifestyle	Ophthalmic Laboratory Technicians (51–9083): share increases as an aging population means more demand for contact lenses and glasses

Sources: This table is based on BLS Employment Projections 2019–2029, Table 1.12 "Factors affecting occupational utilization, projected 2019–29". Reasons are coded based on their own considerations.

https://doi.org/10.1371/journal.pone.0291428.t002

For each of our clusters, we aggregate employment numbers across occupations with a BLS explanation, focusing on those related to technological change. These aggregated values allow us to quantify the share of jobs in each occupational cluster where technological change results in large gains or losses.

Results

Table 8 provides information about the nested nature of the 17 occupational groups within the 8 macro clusters. Appendix D in S2 Appendix contains a detailed list of the occupations included in each of the cluster hierarchies.

Our classification scheme illustrates the degree of human capital fluidity between similar occupations in the contemporary US labor market that is not confined to the boundaries of standard industry classification systems. For example, our Cluster 1 (Managers and Supervisors) includes many occupations that are listed under the "Management Occupations" major group of the SOC 2010 classification (code 11–0000), but it also incorporates occupations from other major groups of the SOC 2010 that are related to management in some specific industry or sphere. For instance, Cluster 1 includes "Directors, Religious Activities and Education," which is listed under "Community and Social Service Occupations" major group in the SOC 2010 (code 21–0000), "Chefs and Head Cooks" which is listed under "Food Preparations and Serving Related Occupations" major group in SOC 2010 (code 35–0000), and "Supervisors of Production Workers" which is listed under "Production Occupations" major group in the SOC 2010 (code 51–0000). All these occupations have relatively similar requirements in terms of characteristics that we use in our analysis and are thus placed in the same cluster in our classification scheme. However, in the SOC 2010 classification scheme, they belong to different major categories or industries. Despite the optimized results derived from PCA clustering analysis, it should be noted that the clustering configuration and labeling may have errors or biases. Prior work has found that it lacks sound validation when performing unsupervised machine learning techniques to identify groups and using a supervised machine learning classifier (e.g., decision tree) to predict the cluster labeling [78].

Cluster characteristics

Tables 9 and 10 contain summaries of the underlying occupational descriptors for 17 occupational groups, with the cluster structure delineated with shading. For educational attainment, experience, and training, O*NET reports the proportion of workers in an occupation that falls within each category: the level ranging from 1 (below high school) to 12 (post-doc) for education and the duration ranging from "0–3 months" to "10+ years" for experience and training. After clustering, we average the proportions of workers within each category of these variables for occupations falling in the same group and use these values as weights to compute the average education level and duration of experience and training for each occupational grouping. For values and interests, O*NET provides an expert assessment of the importance of 6 work values and 6 categories of work interests (ranging from 0–5, with 5 being very important) for a given occupation. These scores are averaged for occupations within each group and for visual convenience are rescaled to be between 0 (corresponding to 0 score) and 10 (corresponding to the highest average group score).

Cluster 1 comprises groups one and two containing management and supervisory occupations. Group 1 includes managers in producer services (e.g., business and financial operations), and Group 2 contains preschool education administrators, chefs and head cooks, residential advisors, and first-line supervisors. These occupations require knowledge in administration and management as well as personnel and human resources. They also require social skills, including negotiation, coordination, persuasion, and complex problem-solving skills. Cluster 2 is comprised primarily of post-secondary education occupations (Group 3) as well as occupations in mental health and counseling (Group 4). These occupations require knowledge in education and training as well as psychology, therapy and counseling. Occupations in Cluster 2 have the highest level of educational attainment: 29% of occupations have a Master's degree, 34% have a Ph.D., and 9% have some sort of post-doctoral experience.

https://doi.org/10.1371/journal.pone.0291428

Table 8. Clustering results based on ward's method.

List of 8 clusters	List of 17 groups	Example occupations
Cluster 1: Managerial and supervisory occupations (n = 71)	Group 1: Managers (n = 43)	Human Resources Managers
		Sales Managers
		Agents/Business Managers of Artists/Athletes
		Transportation Planners
	Group 2: Supervisors (n = 28)	Postmasters and Mail Superintendents
		First-Line Supervisors of Retail Sales Workers
Cluster 2: Education, mental health and counselling occupations (n = 71)	Group 3: Educators (n = 39)	Biological Science Teachers, Postsecondary
		Communications Teachers, Postsecondary
		History Teachers, Postsecondary
	Group 4: Counselors (n = 32)	Clinical, Counseling, and School Psychologists
		Mental Health, Substance Abuse Social Workers
		Special Education Teachers, Secondary School
Cluster 3: Science, technology and engineering occupations (n = 100)	Group 5: Scientists (n = 36)	Biomedical Engineers
		Hydrologists
		Geographers
	Group 6: Non-medical technicians (n = 64)	Private Detectives and Investigators
		Sound Engineering Technicians
		Electrical/Electronic Engineering Technologists
		Agricultural and Food Science Technicians
		Computer Programmers
		Electrical and Electronics Drafters
Cluster 4: Special and support services occupations (n = 76)	Group 7: Medical technicians (n = 19)	Respiratory Therapists
	Group 8: Physicians (n = 36)	Adapted Physical Education Specialists
		Pharmacists
		Registered Nurses
		Oral and Maxillofacial Surgeons
	Group 9: Safety protectors (n = 21)	Police and Sheriff's Patrol Officers
		Air Traffic Controllers
Cluster 5: Analytical and expert services occupations (n = 90)	Group 10: Analysts (n = 37)	Financial Examiners
		Statisticians
		Loan Officers
		Insurance Sales Agents
	Group 11: Clerical workers (n = 22)	Librarians
		Editor
	Group 12: Office staff (n = 31)	Medical Secretaries Data Entry Keyers

(*Continued*)

https://doi.org/10.1371/journal.pone.0291428

Table 8. Clustering results based on ward's method. (*Continued*)

List of 8 clusters	List of 17 groups	Example occupations
Cluster 6: Customer and personal services occupations (n = 86)	Group 13: Personal service providers (n = 55)	Home Health Aids
		Bakers
		Hotel, Motel, and Resort Desk Clerks
		Counter and Rental Clerks
		Cook, Fast Food
	Group 14: Sales persons (n = 31)	Costume Attendants
		Fitness Trainer and Aerobics Instructors
		Retail Sales persons
		Hairdressers, Hairstylists, and Cosmetologists
Cluster 7: Equipment operation and services occupations (n = 84)	Group 15: Equipment operators (n = 84)	Bus and Truck Mechanics, Diesel Engine Specialists
		Elevator Installers and Repairers
		Radio, Cellular, Tower Equipment Installers/ Repairers
		Highway Maintenance Workers
		Crane and Tower Operators
Cluster 8: Basic services, crafts and manual occupations (n = 178)	Group 16: Craftspersons (n = 133)	Watch Repairers
		Fiberglass Laminators and Fabricators
		Industrial Truck and Tractor Operators
		Floor Sanders and Finishers
		Tile and Marble Setters
		Metal-Refining Furnace Operators and Tenders
	Group 17: Manual labors (n = 45)	Locker Room, Coatroom, Dressing Room Attendants
		Laborers and Freight, Stock, Material Movers, Hand
		Light Truck or Delivery Services Drivers

https://doi.org/10.1371/journal.pone.0291428.t003

Cluster 3 contains engineers, computer programmers, technicians, and other occupations where data and analysis are at the core of their daily activities. These occupations require knowledge of the sciences (chemistry, biology and physics), computers and electronics, and engineering and technology. The occupations also require various technical skills, including complex problem-solving, programming, design, troubleshooting, and operations analysis. Cluster 4 contains occupations that require knowledge across a range of domains, including biology, medicine and dentistry, public safety and security and law and government. These occupations require skills in science, complex problem solving and critical thinking, as well as abilities related to problem recognition or problem sensitivity. This cluster has a great deal of variation in educational requirements: 22% of occupations have an associate's degree, 18% have a bachelor's degree, 10% have a Master's degree, 13% have a doctorate, and 8% have postdoctoral education. Occupations in Cluster 5 primarily include analytical, clerical, and office jobs, which deal with data, writing, and paperwork as their core activities. These occupations are similar to Cluster 3 in their skills (e.g., mathematics, critical thinking, complex problem solving) and abilities (e.g., mathematical reasoning, number facility). However, occupations in Cluster 5 involve more structured work than in Cluster 3.

Cluster 6 represents customer and personal services occupations, such as food preparation, retail salespersons, and hairdressers. Cluster 7 groups equipment operation and service occupations, such as bus mechanics, elevator installers, and crane operators. Cluster 8 includes basic services, crafts, and manual occupations, such as watch repairers, material movers, and tile setters. Tables 4 and 5 illustrate that Clusters 6, 7 and 8 are distinct from the other five clusters discussed to this point in time with respect to their levels of educational attainment and work type. Workers in these occupations only require some amount of post-secondary education. In terms of work values, these clusters have a lower average score in achievement, independence, recognition, and work conditions, while having higher average scores on support (as well as relationships in some service-related groups). Work interests in these clusters tend to be relatively more conventional and realistic compared to Clusters 1–5. Cluster 6 is comprised primarily of service occupations that require knowledge about food production and customer and personal service. Clusters 7 and 8 are similar in that they share realistic interests with practical, hands-on work. These occupations often involve working outside with tools and machinery or plants and animals. The majority of occupations in Cluster 7 require mechanical knowledge and technical skills related to equipment maintenance, equipment selection, and repair. Cluster 8 contains occupations that require substantial manual labor and knowledge of the following areas: transportation, production and processing, and customer and personal service. They also require a high degree of technical skills related to equipment selection and maintenance, as well as social skills.

Table 9. The average education, experience and job training.

Cluster	Group	Education level (1–12)	Experience (years)	Job training (years)
C1: Managerial and supervisory occupations	G1: Managers	6.6	4.3	0.9
	G2: Supervisors	4.1	3.1	0.9
C2: Education, mental health and counselling occupations	G3: Educators	9.9	3.6	1.0
	G4: Counselors	7.8	2.2	0.8
C3: Science, technology and engineering occupations	G5: Scientists	7.3	3.4	1.2
	G6: Non-medical technicians	4.8	3.1	1.0
C4: Special and support services occupations	G7: Medical technicians	4.8	1.3	0.6
	G8: Physicians	8.4	2.5	1.0
	G9: Safety protectors	4.2	3.0	1.4
C5: Analytical and expert services occupations	G10: Analysts	5.7	2.7	0.9
	G11: Clerical workers	5.8	2.0	0.6
	G12: Office staff	3.7	1.7	0.6
C6: Customer and personal services occupations	G13: Personal service providers	2.6	1.0	0.5
	G14: Sales persons	2.7	1.3	0.6
C7: Equipment operation and services occupations	G15: Equipment operators	2.6	1.9	1.2
C8: Basic services, crafts and manual occupations	G16: Craftspersons	2.2	1.6	1.1
	G17: Manual labors	2.1	0.8	0.5

Sources: Authors own calculations. Occupations are clustered using the Ward hierarchical method based on 72 principal components derived from 220 occupational descriptors from O*NET. For each occupation, descriptors related to education, experience, and job training are aggregated into 3 corresponding variables and then averaged across occupations in each group.

https://doi.org/10.1371/journal.pone.0291428.t004

Table 10. The average scores for work values and interests.

Cluster	Group	Work Values				Work Values		Work Interests					
		Achievement	Independence	Recognition	Relationships	Support	Working Conditions	Artistic	Conventional	Enterprising	Investigative	Realistic	Social
C1: Managerial and supervisory occupations	G1: Managers	9	9	8	8	7	9	5	6	9	5	4	6
	G2: Supervisors	9	8	8	10	7	8	6	5	5	6	3	10
C2: Education, mental health and counselling occupations	G3: Educators	9	8	8	6	7	8	4	6	4	10	8	3
	G4: Counselors	7	8	7	6	8	7	4	7	4	7	8	3
C3: Science, technology and engineering occupations	G5: Scientists	9	8	8	6	7	8	4	6	4	10	8	3
	G6: Non-medical technicians	7	8	7	6	8	7	4	7	4	7	8	3
C4: Special and support services occupations	G7: Medical technicians	7	7	6	9	9	7	2	6	4	6	8	8
	G8: Physicians	9	9	8	10	7	8	3	4	4	8	7	9
	G9: Safety protectors	8	8	7	8	9	7	2	6	7	5	9	5
C5: Analytical and expert services occupations	G10: Analysts	8	8	7	7	7	7	3	9	8	5	3	4
	G11: Clerical workers	8	7	6	8	6	7	7	5	6	5	3	6
	G12: Office staff	5	6	5	8	7	5	2	10	7	3	4	4
C6: Customer and personal services occupations	G13: Personal service providers	5	6	4	8	7	5	2	8	6	3	8	5
	G14: Sales persons	6	6	5	8	5	5	5	6	7	2	7	5
C7: Equipment operation and services occupations	G15: Equipment operators	5	6	4	6	8	6	2	6	3	4	10	2
C8: Basic services, crafts and manual occupations	G16: Craftspersons	4	5	4	6	7	5	2	6	3	4	10	2
	G17: Manual labors	3	4	3	6	6	4	2	7	4	2	9	3

Sources: Authors own calculations. Occupations are clustered using Ward hierarchical method based on 72 principal components derived from 220 occupational descriptors from O*NET. Descriptors related to Work Values and Work Interests are averaged across occupations in each group and re-scaled to be in the [1–10] interval.

https://doi.org/10.1371/journal.pone.0291428.t005

Wage patterns in clusters

Fig 21 displays information about the mean and standard deviation of the mean annual wage for each of the seventeen groups developed for this study. Appendix E in S2 Appendix contains more detailed information about the mean annual wage, standard deviation, and coefficient of variation for these groups. Groups 1, 3, 5, and 8 have the highest incomes, with a mean wage of at least $80,000. These groups are shaded in green. Examples of highly paid jobs in Group 1 include management occupations with a wage somewhere around $105,000. Group 3 contains primarily education administrators. Group 5 is composed mainly of engineering occupations which are known to be high-paying jobs. The highest mean wage, of about $120,000, is observed in Group 8, which contains some highly paid medical occupations (i.e., orthodontists, surgeons, anesthesiologists). Incidentally, this is also the group with the highest variation relative to the mean (the coefficient of variation in this group reaches 58%). The high coefficient of variation reflects lower wages of some counseling jobs in this group (marriage and family therapists, for example, with an average wage of $55,000 per year) and much higher-paid counseling jobs that require a medical background (for example, psychiatrists earn $178,000 per year).

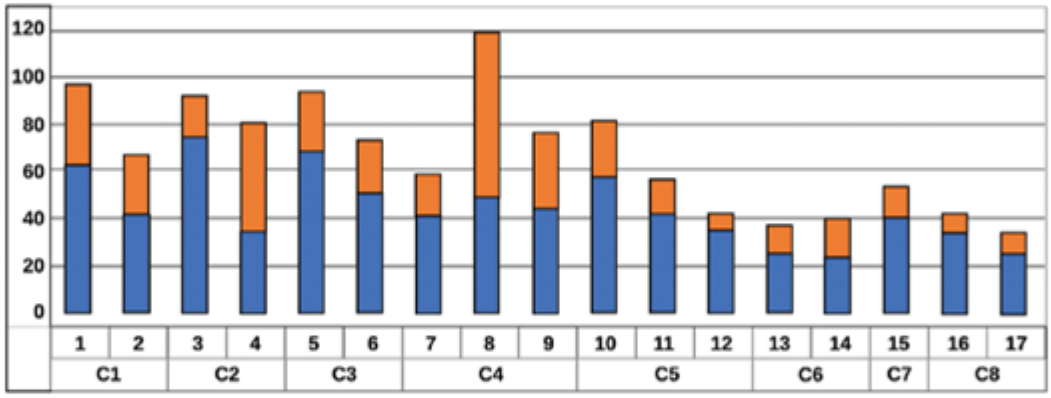


Fig 21. The mean and the standard deviation in wages by occupational groupings. Sources: Based on wage data from the Occupation Employment Statistics survey (BLS), occupations are clustered using the Ward hierarchical method based on 72 principal components derived from 220 occupational descriptors from O*NET. The dots represent the mean wage for each sub-group, and the bars reflect the standard deviations in wages for the given subgroup.

https://doi.org/10.1371/journal.pone.0291428.g002

There are several occupational groups shaded in yellow (Groups 1, 4, 6, 7, 9, 10, 11, 15) that make average wage levels falling somewhere between $50,000 and $80,000. Group two contains a lot of managerial and supervisory positions, while Group 4 contains a lot of teaching, counselors and social workers. Group 9 contains a lot of first responder workers (e.g., paramedics, police officers, firefighters) and occupations working in aviation and nuclear power. Group 15 contains technicians, repairers, and other types of specialized workers (e.g., hazardous materials removal workers).

Groups 12, 13, 14, 16 and 17 contain occupations with below-average wages. Occupations in these groups make between $30,000 and $45,000. Group 12 contains several clerical and secretarial positions, including paralegals and legal assistants. Group 13 contains various low-paid service occupations, including fast food servers, maids and cleaners, cashiers, shipping clerks, and other similar professions. Group 17 contains manual labor occupations (e.g., quarry rock splitters, sewing machine operators, cutters and trimmers and other similar occupations).

Employment trends in clusters

To evaluate the employment trends within occupational groups, we turn to EP data from the Bureau of Labor Statistics. Fig 22 displays the results for the 17-group solution, with detailed statistics provided in Appendix F in S2 Appendix.

There are large differences in cluster sizes based on the number of jobs within them, with Cluster 6 alone, which is the largest cluster, accounting for nearly one-quarter of all jobs in 2019 –as much as the smallest 4 clusters (2, 3, 4 and 7, which account for 5–7% of total employment each). The medium-sized Clusters 1, 5, and 8 contain 15–20% of all jobs each. As discussed above, Clusters 6–8 differ in their human capital requirements and work interest and values, and now we see that together these clusters account for nearly half of all jobs. Group 12 shows many patterns similar to those observed in Clusters 6–8 (by wages and human capital requirements). If we add this group to Groups 13–17 (Clusters 6–8), together, they account for almost 60% of all jobs.

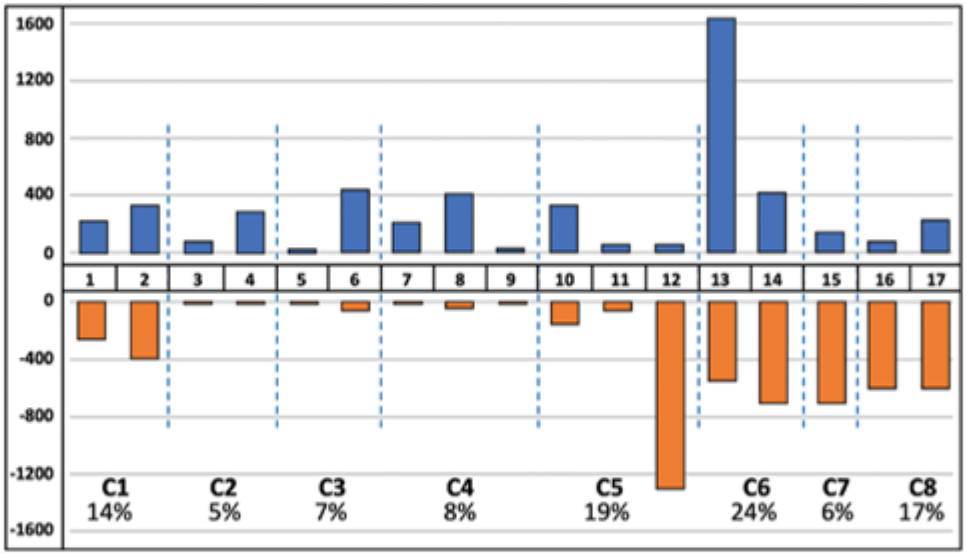


Fig 22. Job losses and gains by occupational groupings (in thousands). Sources: Based on BLS Employment Projections 2019–2029, Table 1.7 "Employment by detailed occupation, 2019 and projected 2029". Occupations are clustered using Ward hierarchical method based on 72 principal components derived from 220 occupational descriptors from O*NET. Gains are shown in blue, and losses–in orange.

https://doi.org/10.1371/journal.pone.0291428.g003

Fig 22 shows that the largest gain in jobs is projected for Cluster 6, especially in Group 13, which has the largest net gain of occupations, despite also having sizable job losses. Most gains in this group are due to expansion in personal assistance-related jobs (e.g., home health aides, physical therapist aids) and jobs related to fine dining (e.g., restaurant cooks, bartenders). Otherwise, job gains are relatively evenly spread across the remaining clusters.

As for job losses, these are mostly concentrated in Groups 1 and 2 (Cluster 1) and in Groups 12–17. Losses in Cluster 1 are more or less compensated with gains of similar size within each group. This indicates that although there will be significant churn or internal mobility within these groups, most changes will not force workers out of the groups, implying relatively low costs of mobility. Instead, in several groups with large job losses, there are no gains of comparable size to accommodate internal mobility, so there will be a push for

workers to move out of their occupational "neighborhood" and face a higher cost of mobility. This is especially pertinent for groups 12 and 16, which have large net losses in jobs. Moreover, the job base is narrowing not only within this group but also in many similar groups, which could further constrain occupational mobility.

Fig 23 demonstrates the role of technology in job dynamics by occupational groups. Appendix G in S2 Appendix provides the supporting statistics. The figure highlights that job losses due to technological change are mostly concentrated in Groups 12–17. Technology accounts for half of all jobs lost in Clusters 6, 7, and 8 and for nearly 90% of jobs lost in Cluster 5 (mostly concentrated in Group 12). Alternatively, Clusters 1–4 appear to be robust to job losses due to technological change. In particular, Clusters 2 and 4, which contain postsecondary education occupations and occupations in science, law, and government, are projected to grow somewhat. This growth is due to other non-technology-related factors (e.g., increased specialization). In Cluster 3, which is comprised of engineers, computer programmers, and technicians, three-quarters of job growth is due to technology.

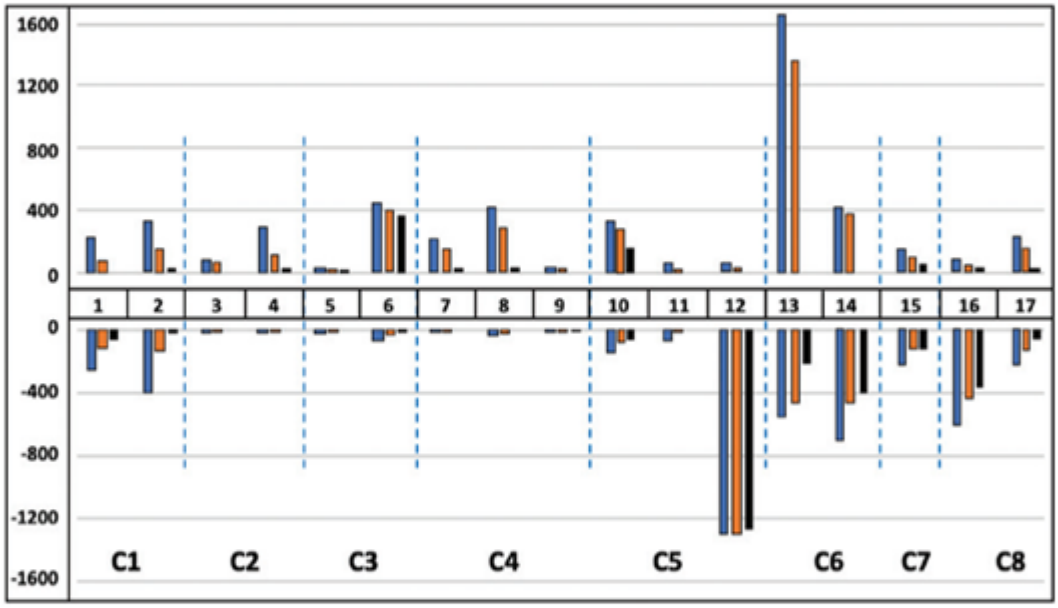


Fig 23. Job losses and gains by occupational groupings, including technological factors. Sources: Based on BLS Employment Projections 2019–2029, Table 1.7 "Employment by detailed occupation, 2019 and projected 2029" and Table 1.12 "Factors affecting occupational utilization, projected 2019–29". Occupations are clustered using the Ward hierarchical method based on 72 principal components derived from 220 occupational descriptors from O*NET. Total change is shown in blue, change that has a reason is shown in orange, and change due to technology is shown in black.

https://doi.org/10.1371/journal.pone.0291428.g004

Fig 24 summarizes three aspects of clusters in terms of their employment trends, educational attainment, and wages. It arranges occupational groupings in a 3x3 matrix where the columns show three categories of education: 1.5 standard deviations below the overall mean of education level ("Below" column including education levels ranging from below high school to post-secondary), 1.5 standard deviations within the overall mean education level ("Average" column, which includes some college, associate degree, BA degree or post-BA certification), and 1.5 standard deviations above the overall mean ("Above" column with education levels ranging from MA diploma to post-doc). The rows are designed in a similar way showing departures from the overall mean annual wage. The net gains and losses in jobs within each group are reflected with

minus/plus symbols, with the number of symbols reflecting the absolute size of the net change. In addition, the column sum at the bottom of this figure indicates the net gain in jobs by education level, and the row sums indicate the net gain in jobs by wage level.

		Education level			Total jobs
		Below	Average	Above	
Annual wage	Below	13 (++++) 14 (--) 16 (---) 17 (·)	12 (------)		-1021
	Average	15 (-)	2 (-) 6 (++) 7 (+) 9 (+) 10 (+) 11 (·)	4 (++)	898
	Above		1 (-)	3 (+) 5 (·) 8 (++)	423
Total jobs		143	-590	747	300

Legend

Negative	Range	Positive
(·)	0-20	(·)
(-)	20-250	(+)
(--)	250-500	(++)
(---)	500-750	(+++)
(----)	750-1000	(++++)
(-----)	1000-1250	(+++++)
(------)	1250 and up	(++++++)

Fig 24. Job losses and gains by occupational groupings—By wage and education level. Sources: Authors calculations based on BLS Employment Projections 2019–2029, Table 1.7 "Employment by detailed occupation, 2019 and projected 2029". Occupations are clustered using the Ward hierarchical method based on 72 principal components derived from 220 occupational descriptors from O*NET.

https://doi.org/10.1371/journal.pone.0291428.g005

A general picture that emerges from examining the row and column totals in Fig 24 suggests the following two processes taking place in the job market. First, there is an expansion in occupations that require below or above-average levels of education, while the demand for jobs described by the average level of education will decrease. Second, occupations that are currently paying low wages are projected to shed many jobs, while occupations with average to high wages are going to expand. This massive migration of jobs into higher paying but relatively distant in terms of human requirements and work type occupational groups could be a costly process. They could require time and assistance from policymakers and human resource specialists. In the short run, it could also mean excess labor in occupational groups with projected job losses, possibly depressing wages paid in these occupations, and an increased probability and duration of unemployment. Instead, in the expanding occupational groups, there could be short-run labor shortages, fueling up already high wages.

A more detailed examination reveals a large portion of gains will come from jobs that have above-average wages and levels of educational attainment (e.g., Groups 3 and 8). Group 3 contains various types of post-secondary educators (e.g., physics, chemistry, political science). Growth in this group likely stems from anticipated increases in undergraduate enrollment at institutions of higher education; by 2029, undergraduate enrollment is projected to increase by 17 million students (NCES, 2020). The educational attainment for both of these occupational groups is substantial. A total of 93% of occupations in Group 3 and 62% of occupations in Group 8 require a Master's degree or higher. Group 8 contains many medical support occupations, such as health educators, pharmacists, and optometrists, as well as surgeons and anesthesiologists. The growth in

this cluster is not surprising, given the anticipated need for medical professionals. As the population ages and people live longer, many will need treatment for chronic health conditions [79].

Job growth is anticipated in occupations that require average levels of educational attainment and pay average wages. Most of the growth in this category is in Group 6, which contains science, technology and media specialists and technicians. A total of 73% of occupations in this group require education beyond high school and up to (but not exceeding) a bachelor's degree. In this group, 80% of the jobs created will be due to technological change. The EP data note the reasons behind this technology connection. For example, more jobs will be needed in cybersecurity to protect sensitive financial and personal data as an increasing amount of information is stored in cloud environments. More jobs will be needed to design, create and implement software to run all the devices connected to the Internet as society moves towards an Internet of Things (IoT). Job growth in Group 10 is also related to the digitization of society, as more people working in occupations that analyze data will be needed to enhance business processes and new products.

Groups with job losses (e.g., 12, 14, 15, 16) are clustered in the lower left cells of Fig 5 and have lower levels of educational attainment and wages. As discussed, these groups will lose a large portion of jobs because of technological change. In fact, for some, such as Group 12, almost all job losses are linked to technological change. Examples of jobs that will be lost include paralegals and legal assistants, new accounts clerks, correspondence clerks, and medical secretaries. These jobs are expected to be replaced by capital as automation enables machines to do the jobs of people or enables fewer workers to perform their jobs more efficiently. Group 16 is projected to lose a large number of jobs due to technological change as well. These job losses are related to capital/labor substitutions. For example, hand packers and packagers are projected to be replaced by machines that will do the same work. Another large chunk of jobs will be lost due to productivity changes. In the future, for example, fewer workers in the job of "Tool Grinders, Filers, and Sharpener" will be needed as programmable equipment will help them to perform their jobs more efficiently.

Three groups that run counter to the trends described above are Groups 1, 2, and 13. In Group 1, which pays above-average wages but requires average levels of educational attainment, technology is expected to replace the following types of jobs: purchasing managers, sales managers, buyers and purchasing agents, and logisticians. This job replacement is projected to be a result of task automation as capital substitutes for labor. These job losses are of concern because of their high wage levels; workers in these occupations may be forced to transition to other occupations that pay less. Group 13 is projected to experience an increase in jobs. Unfortunately, these gains are not necessarily advantageous to workers because of their below-aver-age educational requirements and low wages. Examples of occupations in this group include fast food cooks, restaurant servers, occupational therapy aides, and pharmacy aids. Thus, workers in the occupations may need to work multiple jobs to maintain a particular standard of living.

Discussion

Innovations in technology, including the Internet of Things (IoT), artificial intelligence, autonomous vehicles, and drones, are reshaping the workplace. These innovations are changing the types of work being done and how work is done [80–83]. Technological change has also visibly impacted the labor market as jobs become increasingly bifurcated into low-skill, low-wage jobs and high-skill, high-wage jobs [2, 84, 85]. This bifurcation of work has important implications for the ability of workers to change jobs or occupational mobility.

To this point in time, empirical work related to routine-biased technological change (RBTC) has produced important information about the costs of mobility for older workers [86] and declining wage premiums for routine jobs in manual labor [11]. To our knowledge, studies do not consider the impact of technological change on the size of job clusters prior to analyzing occupational mobility. This a priori step is important

because technological change may eliminate several jobs within occupational clusters, limiting the number of occupational alternatives for workers. Our study addresses this knowledge gap by undertaking a comprehensive clustering of over 700 occupations across multiple dimensions (e.g., KSA, human capital, worker values and interests) to create groups of similar jobs. We presented information on wage variations within occupational clusters to demonstrate their coherence. After summarizing the characteristics of occupational clusters, we analyzed their growth trajectory and assessed the role of technological change on job dynamics. This analysis is insightful for future work assessing occupational mobility since movement between jobs within the same cluster, particularly for finer-resolution cluster solutions, is likely easier than movements between clusters.

Consistent with prior work, analytical results revealed technological change is responsible for a large proportion of job losses in low-wage, low-skill occupations (e.g., Groups 12, 14, 16) [87]. Occupations with job losses had the characteristics of jobs anticipated to be disrupted by technology; routine and somewhat repetitive in nature [11, 12]. Group 12 was a good example of technology replacing workers or enabling businesses to continue operations with fewer workers. The analysis also indicated continued growth of high-wage, high-educational attainment jobs requiring a Master's degree or higher (e.g., post-secondary educators and medical support occupations), as well as high-wage occupations where task automation enables capital to substitute for labor (e.g., purchasing managers, logisticians). Losses in these occupations are less of a concern however because they also require high levels of educational attainment which indicates these workers will be able to find alternative forms of work. The projected large loss of jobs in low-wage groups (e.g., Groups 12, 14, 16) is cause for concern. This will limit alternative jobs that are closest to the lost jobs in terms of educational levels, skills, interests, values etc. and means that workers in these highly impacted groups will likely need to be retrained and/or go back to school in order to move to jobs outside of their occupational group.

The findings also indicated growth of jobs in occupations paying average wages and requiring average levels of educational attainment (an education beyond high school and up to, but not exceeding a college degree). Much of this job growth is related to Internet-related technologies and the Internet of Things (IoT). Moving forward, it will be important to track the wage levels of these high-growth occupational groups to ascertain if wages in these groups decline as more workers fill these positions. If wages do decline, this will certainly indicate a continued trend of disappearing jobs in the middle of the wage spectrum and continued wage inequality in the workforce related to technological change and RBTC [88, 89]. At this point in time, however, our results do not suggest, as RBTC does, that jobs in the middle portion of the wage spectrum will disappear by 2029. This a short-run outcome, however. Wages in these occupations will need to be tracked long-term to ascertain if they grow, decline or stay the same.

Our analytical approach makes several contributions to work on technological change and occupational mobility. First, it provides a comprehensive, multidimensional perspective on occupational grouping, which moves beyond the prior unidimensional emphasis of prior work on either skills, knowledge, tasks, or worker values and interests. Second, our clustering approach created a multi-level taxonomy of similar occupations that provide both macro-level snapshots and detailed micro-level information that may be used in workforce development and training initiatives. Three, our work provides a detailed taxonomy of similar occupations at finer levels of classification (provided in Appendix D in S2 Appendix), which can be of great practical use. For example, human resources specialists and counselors can use this information to advise workers on decisions related to occupational choice and mobility. Four, our analysis assesses the impact of technological change on job dynamics within clusters, which is important for assessing future employment alternatives for workers displaced by technological change.

That said, it is essential to acknowledge some limitations of our work, which relate to assumptions underlying the analysis. Specifically, we assume that the structure of clusters is static, which implies that the characteristics used to cluster occupations will remain unchanged over time. While this assumption is naive if one takes

a long-term perspective, for short-to medium-term analyses, this assumption is not unreasonable. Another assumption is that job gains or losses within occupations where technology was indicated as the primary source of change were fully attributed to technology. In reality, technology might not be the only reason for observed changes. This situation was not very common, however, so we do not expect it to affect our main findings. Our analysis is based on pre-COVID-19 projections, so it is not fully reflective of changes in the labor market that might have been triggered by the effects of the pandemic. Additionally, our analysis does not assess occupational mobility between clusters. Instead, the taxonomy developed in this study may be used to assess movement between occupations in future work. Lastly, people need to be cautioned that the implementation of clustering and interpretation of results may have errors or biases due to the limitation of the algorithm [78, 90]. Interpretation and Implications of the results should be paired with theories and the literature.

From a policy perspective, the taxonomy of occupations and their projected growth provides critical information for policy makers to focus attention on particular segments of the taxonomy projected to experience the greatest pressure from technological change. In this respect, our study provides a foundation for moving forward in a proactive way for worker retraining in several ways. First, our nested occupational clustering provides a foundation for developing worker training and certificate programs by identifying knowledge, skills, and educational gaps between declining and growing occupations. Second, the information about values and interests provides people with a sense of the occupations that are most aligned with their values and interests, which could increase worker satisfaction after moving to alternative occupations.

Another dimension for targeted policymaking involves dedicated attention to groups most vulnerable to technological change. Older workers, for example, might be less likely to invest in acquiring new types of skills and knowledge [91] and might be less adaptable to changes in work due to physical constraints and established habits [92]. Older workers may also have more difficulty adapting to new technology compared to younger workers [93], which means that older workers employed in occupations most affected by technological changes could be facing a higher risk of labor market displacement and higher associated costs. Another example of a vulnerable population is ethnic and racial minorities, who are more likely to be concentrated in occupations associated with low skill requirements and low wages [94]. Racial/ ethnic minorities are also likely to experience more labor market discrimination in times of economic scarcity [95], including the scarcity of jobs in relevant occupational groups.

For vulnerable groups, on-the-job training may be the best means of skills upgrading. For a long time, employers in the United States have not done a good job at providing on-the-job training [61, 62] and studies point to the need for a turning of the tide for displaced workers but also employers who are finding it increasingly difficult to find qualified workers for open positions [96]. On-the-job training may mean hiring workers without college degrees and providing training in skills where companies need workers. This practice would line up with emerging corporate practice, which sees employers moving towards hiring people without four-year college degrees for entry-level positions, particularly for racial/ethnic minorities [97]. Community college programs may also be a means of helping the workforce gain technology-relevant skills in shorter amounts of time, which may be critical as the pace of technological change quickens. Pima Community College in Arizona, for example, has a one-semester certificate program called "Autonomous Vehicle Driver & Operations Specialist," which trains drivers on how to operate and interact with self-driving trucks [98].

Finally, the issue of inequality in educational opportunities has been an increasing concern and exacerbated by the rapid pace of technological advancement, as access to and proficiency with technology has become increasingly important in the workforce [63, 64]. Our analyses of wage patterns and projected employment trends within occupational clusters highlight the urgent need for policy interventions aimed at narrowing the technological gaps in educational opportunities. To effectively address this issue, policymakers must focus on improving access to technology and digital literacy skills for all students throughout childhood and young

adulthood. This can be achieved through initiatives such as providing public schools (especially those in the underrepresented neighborhoods) with more resources for technology infrastructure and training teachers in digital education strategies. Additionally, efforts should be made to increase access to technology outside of the classroom, such as through community educational centers and public libraries.

Conclusions

Technological change will continue to have major effects on the development of the labor market. As technology replaces jobs or changes the nature of work, people may have to transition into new positions. This paper provided new knowledge to the literature on technological change and occupational mobility by classifying occupations across a range of characteristics, including knowledge, skills and abilities (KSA), human capital requirements (education, experience, and training (EET)) and work activities, values and interests (AVI). The framework behind this classification is that movement between jobs within the same cluster is easier than movement between clusters. We then analyzed the anticipated impact of technological change on job growth in these clusters. This analysis revealed that large portions of clusters will be lost because of technology, which poses occupational mobility challenges for workers in these jobs at present. The analysis also revealed losses in low-wage, low-skill jobs and growth in mid-skilled and highly skilled jobs, which suggests a critical need for skills upgrading and workforce development initiatives in the next ten years to prevent job displacement of workers in low-wage occupations. Our findings complement the classic theory of routine-biased technological change, suggesting that many mid-wage, routine-based occupations will continue growing because of technologies in the short run. Special attention should be paid to vulnerable groups of workers concentrated in occupations with jobs most threatened by technological changes, including but not limited to older individuals and representatives of ethnic and racial minorities.

Supporting information

S1 Appendix. Technical appendix.
(DOCX)

S2 Appendix. Appendices.
(DOCX)

Author Contributions

Conceptualization: Karine Torosyan, Elizabeth A. Mack.
Data curation: Karine Torosyan, Jenna A. Van Fossen, Nathan Baker.
Formal analysis: Karine Torosyan.
Funding acquisition: Elizabeth A. Mack.
Investigation: Karine Torosyan, Sicheng Wang, Elizabeth A. Mack.
Methodology: Karine Torosyan, Elizabeth A. Mack.
Project administration: Sicheng Wang.
Software: Sicheng Wang.
Supervision: Elizabeth A. Mack.
Validation: Sicheng Wang.
Visualization: Karine Torosyan, Sicheng Wang.
Writing – original draft: Karine Torosyan, Sicheng Wang, Elizabeth A. Mack, Jenna A. Van Fossen, Nathan Baker.
Writing – review & editing: Sicheng Wang, Elizabeth A. Mack.

References

1. Card D, DiNardo JE. Skill-biased technological change and rising wage inequality: Some problems and puzzles. Journal of labor economics. 2002; 20: 733–783.
2. Goos M, Manning A. Lousy and Lovely Jobs: The Rising Polarization of Work in Britain. The Review of Economics and Statistics. 2007; 89: 118–133. https://doi.org/10.1162/rest.89.1.118
3. Cirillo V, Evangelista R, Guarascio D, Sostero M. Digitalization, routineness and employment: An exploration on Italian task-based data. Research Policy. 2021; 50: 104079.
4. Katz LF, Murphy KM. Changes in relative wages, 1963–1987: supply and demand factors. The quarterly journal of economics. 1992; 107: 35–78.
5. Susskind D. A world without work: Technology, automation and how we should respond. Penguin UK; 2020. https://fred.stlouisfed.org/series/PAYEMS.
6. Zahidi S, Ratcheva V, Hingel G, Brown S. The Future of Jobs Report 2020. World Economic Forum; 2020. https://www.weforum.org/reports/the-future-of-jobs-report-2020/
7. Acemoglu D, Restrepo P. Low-skill and high-skill automation. Journal of Human Capital. 2018; 12: 204– 232.
8. Lloyd C, Payne J. Rethinking country effects: robotics, AI and work futures in Norway and the UK. New Technology, Work and Employment. 2019; 34. Available: https://onlinelibrary.wiley.com/doi/abs/10. 1111/ntwe.12149?casa_token=kDIScybsKP8AAAAA:1C-waXc6gX82vdIMRCeiaJLLfRxRLaxnzsX4W7FF7Th87NMfmBsy0YiFza41DIjU5RzWFJJ1nTkDtA
9. Autor DH, Levy F, Murnane RJ. The Skill Content of Recent Technological Change: An Empirical Exploration*. The Quarterly Journal of Economics. 2003; 118: 1279–1333. https://doi.org/10.1162/003355303322552801
10. Cortes GM. Where Have the Middle-Wage Workers Gone? A Study of Polarization Using Panel Data. Journal of Labor Economics. 2016; 34: 63–105. https://doi.org/10.1086/682289
11. Maczulskij T. Occupational Mobility of Routine Workers. Rochester, NY; 2021.
12. Goos M, Rademakers E, Röttger R. Routine-biased technical change: Individual-level evidence from a plant closure. Research Policy. 2021; 50: 104002.
13. Kerr S, Maczulskij T, Maliranta M. Within and between firm trends in job polarization: the roles of globalization and technology. Journal of Economic Geography. 2020; 20: 1003–1039.
14. Schmutz B, Sidibé M, Vidal-Naquet E. Why are Low-Skilled Workers less Mobile? The Role of Mobility Costs and Spatial Frictions. HAL-SHS. 2020 [cited 14 Sep 2022]. https://halshs.archives-ouvertes.fr/ halshs-02951633
15. Jarvis BF, Song X. Rising intragenerational occupational mobility in the United States, 1969 to 2011. American sociological review. 2017; 82: 568–599. https://doi.org/10.1177/0003122417706391 PMID: 28966346
16. Farooq A. The U-shape of Over-education? Human Capital Dynamics & Occupational Mobility over the Life Cycle. Job Market Papers. 2016 [cited 13 Sep 2022]. https://www.brookings.edu/blog/up/
17. Robinson C. Occupational Mobility, Occupation Distance, and Specific Human Capital. J Human Resources. 2018; 53: 513–551. https://doi.org/10.3368/jhr.53.2.0814-6556R2

18. Kleiner MM, Xu M. Occupational Licensing and Labor Market Fluidity. National Bureau of Economic Research; 2020. https://doi.org/10.3386/w27568

19. Bisello M, Maccarrone V, Fernández-Maías E. Occupational mobility, employment transitions and job quality in Europe: The impact of the Great Recession. Economic and Industrial Democracy. 2022; 43: 585–611.

20. Dlouhy K, Biemann T. Path dependence in occupational careers: Understanding occupational mobility development throughout individuals' careers. Journal of Vocational Behavior. 2018; 104: 86–97.

21. Kosteas VD. Occupational concentration and outcomes for displaced workers—Kosteas—2020—Papers in Regional Science—Wiley Online Library. 2020;99: 977–997.

22. Rinawi M, Backes-Gellner U. Skill prices, skill composition, and the structure of wages | VOCEDplus, the international tertiary education and research database. 2015. https://www.voced.edu.au/content/ ngv:72089

23. Groes F, Kircher P, Manovskii I. The U-shapes of occupational mobility. The Review of Economic Studies. 2015; 82: 659–692.

24. Andrieu E, Jamet S, Marcolin L, Squicciarini M. Occupational transitions: The cost of moving to a "safe haven." Paris: OECD; 2019 Feb. https://doi.org/10.1787/6d3f9bff-en

25. Poletaev M, Robinson C. Human Capital Specificity: Evidence from the Dictionary of Occupational Titles and Displaced Worker Surveys, 1984–2000 | Journal of Labor Economics: Vol 26, No 3. Journal of Labor Economics. 2008;26. Available: https://www.journals.uchicago.edu/doi/full/10.1086/588180?casa_token=vt_uUoW1ksMAAAAA%3Aup5y4LnoCT81j8iQTMcp6ZDDAmZFHEHqsRgnrc8UIUuRs kbgURYCssVxdqR2DcqAMnq6SWuCAA

26. Gathmann C, Schönberg U. How general is human capital? A task-based approach. Journal of Labor Economics. 2010; 28: 1–49.

27. Betsworth DG, Fouad NA. Vocational interests: A look at the past 70 years and a glance at the future. The Career Development Quarterly. 1997; 46: 23–47.

28. Hansen J-IC, Wiernik BM. Work Preferences: Vocational Interests and Values. The SAGE Handbook of Industrial, Work and Organizational Psychology. 55 City Road: SAGE Publications Ltd; 2018. pp. 408– 445.

29. Yang Y, Li J, Wu X, Wang J, Li W, Zhu Y, et al. Factors influencing subspecialty choice among medical students: a systematic review and meta-analysis | BMJ Open. BMJ Open. 2019 [cited 13 Sep 2022]. Available: https://bmjopen.bmj.com/content/9/3/e022097.abstract

30. Forsythe E. Careers within firms: Occupational mobility over the lifecycle. Labour. 2019; 33: 241–277.

31. Yamaguchi S. Tasks and Heterogeneous Human Capital. Journal of Labor Economics. 2012; 30: 1–53. https://doi.org/10.1086/662066

32. Bureau of Labor Statistics. Revising the Standard Occupational Classification System. 1999.

33. Bureau of Labor Statistics. Revising the Standard Occupational Classification. 2014.

34. O*NET. About O*NET at O*NET Resource Center. [cited 3 May 2023]. https://www.onetcenter.org/ overview.html

35. Feser EJ. What regions do rather than make: A proposed set of knowledge-based occupation clusters. Urban Studies. 2003; 40: 1937–1958.

36. Koo J. How to Analyze the Regional Economy With Occupation Data. Economic Development Quarterly. 2005; 19: 356–372. https://doi.org/10.1177/0891242405279910

37. Nolan C, Morrison E, Kumar I, Galloway H, Cordes S. Linking Industry and Occupation Clusters in Regional Economic Development. Economic Development Quarterly. 2011; 25: 26–35. https://doi.org/ 10.1177/0891242410386781

38. Slaper TF. Clustering Occupations. Indiana Business Review; 2014. Available: https://www.ibrc.indiana.edu/ibr/2014/summer/article2.html

39. Burrus J, Jackson T, Xi N, Steinberg J. Identifying the Most Important 21st Century Workforce Competency: An Analysis of the Occupational Information Network (O*NET). 2013 Dec pp. i–55. Available: https://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2013.tb02328.x

40. Manzella J, Totty E, Benedetto G. A Task-based Approach to Constructing Occupational Categories with Implications for Empirical Research in Labor Economics. US Census Bureau, Center for Economic Studies; 2019.

41. Yu J-C, Sokolowski HM, Rao KS, Moraglia LE, Khoubrouy SA, Abdi H, et al. Visualization of latent components assessed in O*Net occupations (VOLCANO): A robust method for standardized conversion of occupational labels to ratio scale format. Behav Res. 2023 [cited 18 Apr 2023]. https://doi.org/10.3758/ s13428-022-02044-7 PMID: 36698000

42. Waters K, Shutters ST. Skills-approximate occupations: using networks to guide jobs retraining. Appl Netw Sci. 2022; 7: 1–13. https://doi.org/10.1007/s41109-022-00487-7 PMID: 35789701

43. Shutters ST, Waters K. Inferring Networks of Interdependent Labor Skills to Illuminate Urban Economic Structure. Entropy. 2020; 22: 1078. https://doi.org/10.3390/e22101078 PMID: 33286847

44. Alabdulkareem A, Frank MR, Sun L, AlShebli B, Hidalgo C, Rahwan I. Unpacking the polarization of workplace skills. Science Advances. 2018; 4. https://doi.org/10.1126/sciadv.aao6030 PMID: 30035214

45. Farinha T, Balland P-A, Morrison A, Boschma R. What drives the geography of jobs in the US? Unpacking relatedness. Industry and Innovation. 2019; 26: 988–1022. https://doi.org/10.1080/13662716.2019. 1591940

46. Becker GS. Human capital: A theoretical and empirical analysis, with special reference to education. University of Chicago press; 2009.

47. Becker GS. Human capital revisited. Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education, Third Edition. The University of Chicago Press; 1994. pp. 15–28.

48. Lysova EI, Allan BA, Dik BJ, Duffy RD, Steger MF. Fostering meaningful work in organizations: A multilevel review and integration. Journal of Vocational Behavior. 2019; 110: 374–389. https://doi.org/10. 1016/j.jvb.2018.07.004

49. Gheaus A, Herzog L. The Goods of Work (Other Than Money!). Journal of Social Philosophy. 2016; 47: 70–89. https://doi.org/10.1111/josp.12140

50. Holland JL. Making vocational choices: A theory of vocational personalities and work environments. Psychological Assessment Resources; 1997.

51. Armstrong PI, Day SX, McVay JP, Rounds J. Holland's RIASEC model as an integrative framework for individual differences. Journal of Counseling Psychology. 2008; 55: 1–18. https://doi.org/10.1037/0022-0167.55.1.1

52. Nye CD. Assessing interests in the twenty-first-century workforce: building on a century of interest measurement. Annual Review of Organizational Psychology and Organizational Behavior. 2022; 9: 415– 440.

53. Hanna A, Rounds J. How accurate are interest inventories? A quantitative review of career choice hit rates. Psychological Bulletin. 2020; 146: 765–796. https://doi.org/10.1037/bul0000269 PMID: 32686949

54. Arieli S, Sagiv L, Roccas S. Values at Work: The Impact of Personal Values in Organisations. Applied Psychology. 2020; 69: 230–275. https://doi.org/10.1111/apps.12181

55. Schwartz SH. Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. In: Zanna MP, editor. Advances in Experimental Social Psychology. Academic Press; 1992. pp. 1–65.

56. Edwards JR, Shipp AJ. The relationship between person-environment fit and outcomes: An integrative theoretical framework. Perspectives on organizational fit. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers; 2007. pp. 209–258.

57. Swanson JL, Schneider M. Minnesota theory of work adjustment. Career development and counseling: Putting theory and research to work. 2013; 29–53.

58. Rounds J, Su R. The Nature and Power of Interests. Curr Dir Psychol Sci. 2014; 23: 98–103. https://doi. org/10.1177/0963721414522812

59. Strong EK, Campbell DP, Hansen J-IC. Strong-Campbell Interest Inventory of the Strong Vocational Interest Blank. Stanford University Press; 1981.

60. Lewis P, Rivkin D. Development of the O* NET interest profiler. Raleigh, NC: National Center for O* NET Development. 1999.

61. Fadulu L. Why Is the U.S. So Bad at Worker Retraining? In: The Atlantic [Internet]. 4 Jan 2018 [cited 13 Sep 2022]. https://www.theatlantic.com/education/archive/2018/01/why-is-the-us-so-bad-at-protecting-workers-from-automation/549185/

62. Preston C. The Recession Exposes the US' Failures on Worker Retraining | WIRED. 2021 [cited 13 Sep 2022]. https://www.wired.com/story/recession-exposes-us-failures-worker-retraining/

63. Lee J-W, Lee H. Human capital and income inequality*. Journal of the Asia Pacific Economy. 2018; 23: 554–583. https://doi.org/10.1080/13547860.2018.1515002

64. Ozer M, Perc M. Dreams and realities of school tracking and vocational education. Palgrave Commun. 2020; 6: 1–7. https://doi.org/10.1057/s41599-020-0409-4

65. US Department of Labor. O*NET. In: DOL [Internet]. [cited 30 Apr 2023]. http://www.dol.gov/agencies/ eta/onet

66. O*NET Resource Center. O*NET Database. [cited 13 Sep 2022]. https://www.onetcenter.org/database.html#individual-files

67. U.S. Bureau of Labor Statistics. Employment Projections Home Page: U.S. Bureau of Labor Statistics. [cited 13 Sep 2022]. https://www.bls.gov/emp/

68. Bureau of Labor Statistics. Handbook of Methods-Employment Projections. 2021 [cited 3 May 2023]. https://www.bls.gov/opub/hom/emp/

69. U.S. Bureau of Labor Statistics. EP Data Tables: U.S. Bureau of Labor Statistics. 2021 [cited 13 Sep 2022]. https://www.bls.gov/emp/tables.htm

70. Lance GN, Williams WT. A general theory of classificatory sorting strategies: 1. Hierarchical systems. The computer journal. 1967; 9: 373–380.

71. Jain AK, Dubes RC. Algorithms for clustering data. Prentice-Hall, Inc.; 1988.

72. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons; 2009.

73. Everitt BS, Landau S, Leese M, Stahl D. Cluster analysis 5th ed. John Wiley; 2011.

74. Duda RO, Hart PE, Stork DG. Pattern classification and scene analysis. Wiley New York; 1973.

75. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. Psychometrika. 1985; 50: 159–179.

76. Autor DH, Handel MJ. Putting tasks to the test: Human capital, job tasks, and wages. Journal of labor Economics. 2013; 31: S59–S96.

77. FRED Economic Data. All Employees, Total Nonfarm. In: FRED, Federal Reserve Bank of St. Louis [Internet]. FRED, Federal Reserve Bank of St. Louis; [cited 13 Sep 2022]. https://fred.stlouisfed.org/ series/PAYEMS

78. Rodríguez Andrés A, Amavilah VHS, Otero A. Evaluation of technology clubs by clustering: a cautionary note. Applied Economics. 2021; 53: 5989–6001. https://doi.org/10.1080/00036846.2021.1934393

79. Hannon K. 5 Health Care Jobs on the Rise—The New York Times. In: The New York Times [Internet]. 2018 [cited 13 Sep 2022]. https://www.nytimes.com/2021/04/14/health/health-care-jobs.html

80. Domini G, Grazzi M, Moschella D, Treibich T. Threats and opportunities in the digital era: automation spikes and employment dynamics. Research Policy. 2021; 50: 104137.

81. Lo Turco A, Maggioni D. The knowledge and skill content of production complexity. Research Policy. 2020; 104059. https://doi.org/10.1016/j.respol.2020.104059

82. Holm JR, Lorenz E, Nielsen P. Work organization and job polarization. Research Policy. 2020; 49: 104015.

83. Falck O, Heimisch-Roecker A, Wiederhold S. Returns to ICT skills. Research policy. 2021; 50: 104064.

84. Autor D. Work of the Past, Work of the Future. National Bureau of Economic Research; 2019.

85. Harrigan J, Reshef A, Toubal F. The march of the techies: Job polarization within and between firms. Research Policy. 2021; 50: 104008.

86. Autor DH, Dorn D. The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market. American Economic Review. 2013; 103: 1553–1597. https://doi.org/10.1257/aer.103.5.1553

87. Frey CB, Osborne MA. The future of employment: How susceptible are jobs to computerisation? Technological Forecasting and Social Change. 2017; 114: 254–280. https://doi.org/10.1016/j.techfore.2016. 08.019

88. Goos M, Manning A, Salomons A. Explaining Job Polarization: Routine-Biased Technological Change and Offshoring. American Economic Review. 2014; 104: 2509–2526. https://doi.org/10.1257/aer.104.8. 2509

89. Goos M. The impact of technological progress on labour markets: policy challenges. Oxford review of economic policy. 2018; 34: 362–375.

90. Faber NM, Meinders MJ, Geladi P, Sjöström M, Buydens LMC, Kateman G. Random error bias in principal component analysis. Part I. derivation of theoretical predictions. Analytica Chimica Acta. 1995; 304: 257–271. https://doi.org/10.1016/0003-2670(94)00585-A

91. Taylor P, Urwin P. Age and Participation in Vocational Education and Training. Work, Employment and Society. 2001; 15: 763–779. https://doi.org/10.1177/095001701400438198

92. Hansson RO, Robson SM, Limas MJ. Stress and coping among older workers. Work. 2001; 17: 247–256. PMID: 12441603

93. Friedberg L. The Impact of Technological Change on Older Workers: Evidence from Data on Computer Use. ILR Review. 2003; 56: 511–529. https://doi.org/10.1177/001979390305600309

94. Byars-Winston A, Fouad N, Wen Y. Race/ethnicity and sex in US occupations, 1970–2010: Implications for research, practice, and policy. Journal of vocational behavior. 2015; 87: 54–70.

95. Krosch AR, Tyler TR, Amodio DM. Race and recession: Effects of economic scarcity on racial discrimination. Journal of personality and social psychology. 2017; 113: 892. https://doi.org/10.1037/ pspi0000112 PMID: 28910122

96. Casey M. Looking towards the future: Automation, training, and the middle class. In: Brookings [Internet]. 22 Jul 2020 [cited 13 Sep 2022]. https://www.brookings.edu/blog/up-front/2020/07/22/looking-towards-the-future-automation-training-and-the-middle-class/

97. Cutter C. Some CEOs Suggest Dropping Degree Requirements in Hiring—WSJ. 2021 [cited 13 Sep 2022]. https://www.wsj.com/articles/some-ceos-suggest-dropping-degree-requirements-in-hiring-11620233566

98. Pima Community College. Autonomous Vehicle Driver & Operations Specialist | Pima Community College, Tucson, Arizona. [cited 13 Sep 2022]. https://www.pima.edu/academics-programs/degrees-certificates/industry-manuf-construction/truck-driving/autonomous-vehicle-specialist-cert/index.html

7.9 Discussion Topics

The main theme for this chapter's discussion is about the Evaluating & Controlling Technology. Here are some suggested discussion points:

1. The Importance of Ethical Evaluation in the Development of Technology
   - Is "ethics by design" a realistic standard or an idealistic goal in tech development?
   - Should ethical evaluation be mandated in the tech industry—just like environmental impact assessments?
   - Who should be responsible for ethical oversight—engineers, ethicists, governments, or the public?
   - Can innovation and ethics coexist, or do they inherently conflict in competitive industries?
   - How can ethical blind spots be avoided when technologies are created by culturally homogeneous teams?
2. Strategies for Assessing the Impact of Technology on Society
   - What methods (e.g., social impact assessments, scenario planning) are most effective in predicting long-term tech consequences?
   - Are current strategies reactive or proactive in assessing technology's societal effects?
   - Should citizen panels or public forums be included in evaluating technological impact?
   - Can quantitative data (e.g., usage stats, economic growth) adequately capture social impacts, or do we need more qualitative approaches?
   - How can unintended consequences be identified before technologies are widely deployed?
3. The Neo-Luddite Critique of the Digital Age
   - Are Neo-Luddite concerns about technological overreach and dehumanization valid in the age of AI and surveillance?
   - Does rejecting or resisting certain technologies automatically equate to being "anti-progress"?
   - What can modern technologists learn from Neo-Luddite critiques about privacy, autonomy, and community?
   - Can Neo-Luddite perspectives serve as a necessary counterbalance to unchecked innovation?
   - Where is the line between legitimate critique and fear of change when it comes to new technology?
4. The Role of Stakeholders in Regulating Technology
   - How should power be balanced between government, industry, and civil society in tech regulation?
   - What role should marginalized or underrepresented communities play in shaping tech policies?
   - Are current tech regulations sufficient, or do they lag behind innovation?
   - How can international cooperation be fostered for global tech challenges (e.g., AI, data privacy, bioethics)?
   - Should companies be self-regulating, or is independent oversight necessary to prevent ethical lapses?

5. The Ethical Implications of Emerging Technologies, such as AI and Biotechnology
    - Should AI systems have ethical reasoning capabilities—or should only humans be responsible for their outcomes?
    - Where do we draw ethical boundaries in biotechnology—e.g., in gene editing, cloning, or synthetic biology?
    - How do we prevent algorithmic discrimination and ensure fairness in AI decision-making?
    - Do emerging technologies risk widening inequality, and how can we ethically address this?
    - Is it ethical to use biotechnology to enhance human traits (e.g., intelligence, strength), or should we limit use to therapeutic purposes only?

Make use of the knowledge you gain from this book and your own experience, opinion and judgment will make your post more attractive for discussion.

7.10 Projects

7.10.1 Discerning Truth in the Digital Age: Challenges and Solutions for Information Literacy

Abstract: The research project consists of four weeks that evaluates the public's reported differences in detecting dependable versus misleading information through online searches. This research investigates both the severity of false information problems and identifies methods to combat these problems in essential domains such as health and financial areas. The project evaluates digital literacy through current assessment of the situation to determine problem magnitude and implement solutions that boost public online information credibility assessment capabilities.

Research Questions: The public's online information reliability detection ability stands at what level of importance? The public's dependency on incorrect information in vital fields such as health and finance generates which types of outcomes? The general public needs which measures to boost their digital literacy so they can properly assess web-based information.

Methodology: The research involves reviewing existing studies about information literacy alongside the spread of misinformation that occurs through online sources. Surveys: Create and distribute questionnaires to different populations to determine their capabilities in source verification and their knowledge about misinformation. The analysis focuses on educational programs along with digital literacy initiatives to assess their effectiveness. Statistical Analysis: A statistical evaluation of survey responses helps establish relationships between participant demographics and their performance at information quality assessment.

Expected Outcomes: The report establishes how severe the problem is concerning public online credibility identification abilities. The report summarizes public information discernment skill weaknesses identified through survey results. This project recommends both educational programs and public policies that will enhance information literacy skills among the general public. A collection of proposed tools along with resources exists to help people assess the reliability of online information.

Timeline:

- Week 1: The first week of the project involves its initiation as well as conducting a literature review. Define project scope and research methodology. The initial phase of this project requires researchers to conduct an extensive review of digital literacy research.
- Week 2: Survey Design and Distribution. The survey instrument needs development to evaluate how the public determines reliable information from unreliable sources. The survey reaches different online platforms to gather responses from multiple demographics.
- Week 3: Data Collection and Analysis. The research team gathers survey responses before starting an initial data analysis process. The study examines various established interventions along with their success rates in digital literacy development.
- Week 4: Synthesis and Final Reporting. The research team performs complete data analysis while integrating literature review insights into the study. The writer creates a final report that contains analysis alongside findings and recommendations. A presentation or infographic should be created to present project results in a visual format.

Conclusion: Through this project students will enhance the comprehension of digital information environment obstacles that affect the public and offer practical solutions to boost information literacy which leads to a more discerning and informed society.

7.10.2 Navigating the Web of Information: Assessing and Enhancing Digital Information Literacy

Abstract: Online digital information dominates the environment to such an extent that people face major difficulties in distinguishing trustworthy sources from fake ones. This four-week undergraduate project will determine the extent to which the public struggles to tell reliable from fake information on the internet especially for important topics like health and finance. The research will also consider solutions that may be implemented to improve the general public's information literacy. This project evaluates the existing information environment of misdirection and assesses digital literacy solutions to create practical educational approaches that reduce digital misinformation threats.

Research Questions: How widespread is the problem of web users being unable to find trustworthy information and what results from this inability? What age range and education level and how people use the web influence their ability to identify trustworthy online sources? What educational approaches should we develop or enhance to boost the digital information literacy of the general population?

Methodology: Literature Review: Examine studies on digital literacy, misinformation distribution, and necessary skills to evaluate information. Surveys and Interviews: A survey of different age groups will help determine how well people can detect reliable information on the internet. Add detailed interviews to discover individual circumstances and problems faced by the participants. Educational Research: A review of present-day educational programs and digital literacy initiatives will be done to determine their breadth, extent, and effectiveness. Data Analysis: Qualitative and quantitative data analysis methods will be used to analyze survey and interview results and to determine the impact of educational programs.

Expected Outcomes: A detailed report will describe the degree of the problem concerning digital information literacy. Research findings that show the main demographic characteristics that influence information discernment skills. Actionable recommendations for educational interventions and policy measures to enhance digital literacy. Suggested further research to identify additional ways of dealing with misinformation in the digital world.

Timeline:

- Week 1: Introduction and Literature Review. Gain a clear understanding of the project goals and parameters. Begin to read academic articles, reports, and relevant case studies about digital literacy and misinformation.
- Week 2: Survey Design and Distribution. A survey must be created to determine how well the public can identify trustworthy online information. The survey should be distributed as widely as possible to get a representative sample.
- Week 3: Interviews and Educational Program Review. Interviews with survey participants will provide more detailed information regarding their survey answers. Examine and evaluate the existing digital literacy programs.
- Week 4: Data Analysis and Report Compilation. The collected survey and interview data will be analyzed. Summarize research results in a final report which will be based on the analysis and include recommendations. Create an executive summary and presentation that highlights the most important aspects of the project.

Conclusion: The students will help shape the vital discussion about digital literacy by showing how the general public falls victim to misinformation and by suggesting ways for the public to use the internet more critically.

7.11 Exercises

1. Why is ethical evaluation considered essential in the development of modern technologies?
2. How can ethical evaluation help prevent bias and discrimination in technologies like AI?
3. What role does ethical evaluation play in minimizing unintended consequences of new technologies?
4. How does ethical evaluation support environmental sustainability in the tech industry?
5. In what ways does ethical evaluation promote public trust and social inclusion in technology?
6. Why is a multidisciplinary approach important in assessing the societal impact of technology?
7. How do ethical frameworks contribute to technology impact assessment?
8. What is the purpose of scenario analysis and modeling in evaluating new technologies?
9. In what ways do public engagement and participatory approaches improve technology assessment?
10. Why is continuous monitoring and evaluation necessary after a technology is deployed?
11. What is the main goal of the Neo-Luddite critique of modern technology?
12. Why are Neo-Luddites concerned about privacy in the digital age?
13. How do Neo-Luddites view the impact of automation and artificial intelligence on employment?
14. What environmental concerns do Neo-Luddites raise about modern technology?
15. Do Neo-Luddites oppose all technology? Why or why not?
16. Why is stakeholder collaboration important in setting ethical standards for technology?
17. How do stakeholders help address the gap between technological advancement and regulation?
18. What is the benefit of including diverse stakeholder perspectives in regulation?
19. How do stakeholders contribute to public trust in technology regulation?
20. Why is international stakeholder collaboration crucial in the digital age?
21. Why is bias in AI decision-making considered an ethical concern?
22. What ethical dilemma arises from AI systems like autonomous cars in accident scenarios?
23. What are the ethical risks associated with genetic modification in biotechnology?
24. How does AI-powered surveillance pose ethical challenges?
25. What is the ethical responsibility regarding biotechnology's impact on the environment?

Chapter

8

Professional Ethics and Responsibilities

- The ethical responsibilities of professionals working in the technology industry
- Examining Contrasting Ethical Perspectives
- Examples of ethical violations and their impact on society
- The role of professional organizations in promoting ethical behavior
- Strategies for maintaining ethical standards in the technology industry
- Whistleblowing: Exposing Wrongdoing vs. Honor in Professionalism

8.1 The Ethical Responsibilities of Professionals Working in the Technology Industry

Technology has become an essential part of modern life as it controls how people communicate, work and live their lives. People working in this field are responsible for pushing innovation while also having the important duty to maintain ethical standards. The quick pace of technological development has created numerous ethical challenges that include data privacy issues and AI discrimination along with automation effects on employment and broader- societal consequences. This session examines the ethical duties of technology professionals alongside the extended results of their professional choices [89, 149, 178, 177, 212, 179].

1. Upkeeping data privacy and security measures

   The new age economy functions on data as its primary resource which makes it essential for technology experts to maintain absolute protection of user privacy. The extensive practice of personal information collection by tech companies has led to worries about surveillance capitalism and information misuse. Technology professionals must establish strong security systems and clear data handling practices and obtain user consent to maintain proper handling of personal information.

2. Algorithmic Bias and Fairness Mitigating Algorithmic Bias and Fairness

   Decision-making processes increasingly rely on artificial intelligence (AI) and machine learning algorithms for both criminal justice and hiring functions. The algorithms tend to maintain biases which exist in training data thus resulting in discriminatory outcomes. The ongoing effort of professionals should focus on bias detection followed by algorithm improvement to guarantee fair treatment for every individual regardless of their background.

3. Promoting Transparency and Accountability

   The intricate technology systems create products and services that remain difficult for regular users to understand. The ethical standards require professionals to guarantee transparency through straightforward explanations of their technology operations and data collection processes and their possible impacts on users. The industry must develop accountability systems which help staff members both admit mistakes and use them as learning opportunities.

4. Navigating the Social and Environmental Impact

   Technological advancements in the sector strongly impact both social structures and environmental systems. The creation of professionals should lead them to evaluate the complete effects of their inventions since these may intensify social gaps or damage the environment. The development of responsible technology requires implementing methods which both achieve beneficial social transformations and reduce negative environmental repercussions.

5. Addressing Automation and Job Displacement

   Automation along with artificial intelligence has produced worries about job reduction and employment decline. The tech industry professionals must analyze the societal disturbances generated by new technologies while developing solutions that unite operational improvements with human employment protection. The solution could take the form of reskilling programs or support for affected communities as well as rethinking how work functions in automated environments.

6. Collaborating with Regulatory Authorities

   The ethical duty of professionals includes working with regulatory bodies to verify that technological advancements remain within legal and ethical boundaries. The professionals should participate in policy discussions by providing their expertise to develop safety-focused guidelines which defend both personal interests and societal welfare.

In conclusion, the extensive ethical duties faced by technology professionals need absolute dedication to both technological advancement and moral conduct. Maintaining a proper equilibrium between technological innovation and user protection and societal well-being remains a continuous obstacle. Through data privacy maintenance together with algorithmic bias reduction and transparency promotion and social and environmental impact assessment and job displacement mitigation and regulatory collaboration professionals can build an advanced ethical technological framework. Their work creates a future where technology functions as a transformative power for good.

8.2 Examining Contrasting Ethical Perspectives

The human ethical landscape presents complex situations which ethical theories guide through their frameworks to understand and address. Different ethical frameworks create distinct approaches to evaluate ethical actions because they use distinct foundational principles and practical frameworks. The session introduces multiple ethical perspectives that demonstrate their main differences.

According to Consequentialism particularly Utilitarianism the ethical value of any action derives from its final results. When an action leads to the highest benefits for the maximum number of people it becomes the most ethical choice. The philosophical movement of Utilitarianism founded by Jeremy Bentham and John Stuart Mill functions as a consequentialist system to evaluate actions through their resulting happiness and utility. The core characteristic of this philosophy emerges from its quantitative method which Bentham described as "felicific calculus" to measure pleasure and pain quantities.

Deontological ethics rests on the understanding that moral acts exist independently from their resulting effects. Immanuel Kant defended the idea that moral conduct emerges from duty which follows universal moral principles that he called categorical imperatives. The core principle of Kant's theory depends on the requirement to base all actions on principles which could serve as general laws for everyone.

According to Virtue Ethics the main focus should be on the moral character of the agent rather than their actions or outcomes. Virtues serve as the foundation of this philosophical approach since they represent character traits which lead people toward success and their life's purpose. The central element of virtue ethics stands in its concentration on ethical character growth instead of moral rules or actions.

According to Care Ethics interpersonal relationships gain priority over all else while virtues including care and compassion and empathy take center stage. The theory opposes conventional ethical principles by supporting an examination of context alongside human relationships in ethical decision processes. The care ethics theory stands out because it highlights the essential role of relationships in human life while emphasizing that caring represents a vital life component according to Carol Gilligan and Nel Noddings.

The foundation of Contractarianism consists of moral norms emerging from concepts related to social agreements. The ethical theory of Contractarianism finds its roots in the ideas of Thomas Hobbes and John Rawls. According to this philosophy people agree either explicitly or by implication to create rules which provide mutual advantages. This view stands out through its reliance on rational consent since morality emerges from what rational beings would decide to agree upon through social contracts.

Existentialist Ethics with its influence from Søren Kierkegaard and Jean-Paul Sartre teaches that human beings maintain freedom and bear complete responsibility for their decisions. Existentialist philosophy rejects external moral regulations because people need to create their personal values through genuine decision-making. Existentialist ethics remains distinctive because it centers on three main principles: freedom, authenticity and personal value subjectivity.

According to Relativism ethical standards lack universal applicability because ethical values depend on cultural or individual factors or societal contexts. The theory of Moral relativism questions the existence of absolute moral rules by showing that ethical values differ between different circumstances. Moral relativism distinguishes itself from moral absolutism by defending multiple ethical systems.

The different ethical perspectives provide distinct understandings about what morality consists of. The consequentialist and deontological approaches base their ethical frameworks on actions but differ in their focus on results versus duties while virtue ethics focuses on the development of moral character. The moral significance of relationships stands at the heart of Care ethics while contractarianism establishes its foundation on social agreements between people. Existentialist ethics honors both personal freedom and authentic conduct while relativism acknowledges the diverse moral standards which exist between different contexts.

In conclusion, the comprehension of these differences leads to proper engagement with complex moral situations. Recognizing multiple ethical perspectives allows individuals to better explore the moral realm while evaluating positions and developing complete solutions for ethical problems.

8.3 Examples of Ethical Violations and Their Impact on Society

Ethical violations represent acts that break social and moral rules which create significant effects on people and their communities and the entire society. Various fields including business and healthcare and technology and politics witness ethical misconduct which produces trust destruction and public disturbances and long-term system degradation. During this session we will review selected ethical violations and their major social consequences.

1. Enron Scandal: Corporate Deception and Financial Ruin

   The Enron scandal of the early 2000s serves as a quintessential example of corporate ethical violations. Executives of the company carried out deceptive accounting methods which produced false profits while hiding financial liabilities until bankruptcy occurred. The consequences of the scandal caused thousands of employees to lose their jobs while simultaneously damaging financial institution trust and leading to new standards for corporate governance and accounting.

2. Volkswagen Emissions Scandal: Environmental Deception

   The emissions scandal at Volkswagen demonstrates how ethical violations result in environmental destruction and harm to public health. Volkswagen implemented software which misreported emissions test results thus generating excessive environmental pollution. The scandal damaged

Volkswagen's reputation while sparking discussions about emissions regulation standards and automobile industry integrity as well as the need for immediate clean energy transitions.

3. Cambridge Analytica: Data Privacy Breach and Manipulation

   The Cambridge Analytica scandal exposed how political campaigns exploit personal data through improper means. The company gained Facebook user information through unauthorized means to develop political advertisements which influenced voter decisions. Data privacy violations caused public worries about democratic process manipulation alongside the need for enhanced data protection laws and ethical concerns regarding advertising operations that use data.

4. Opioid Crisis: Medical Ethics and Public Health

   The opioid crisis in the United States developed from multiple factors including medical ethics issues and pharmaceutical industry actions along with inadequate regulatory oversight. Pharmaceutical companies pushed opioid painkillers onto the market by minimizing their addictive potential which caused widespread addiction that resulted in numerous deaths. The crisis demonstrates the dangerous outcomes when financial gains surpass patient care while exposing healthcare providers and drug manufacturers to ethical accountability.

5. Social Media Manipulation: Spreading Misinformation

   Free speech ethical concerns together with technology company content curation responsibilities have emerged because of the growing spread of false information and hateful content across social media platforms. These platforms have been used to create social divisions which have reduced media credibility and intensified public discourse polarization. The solution to this challenge demands finding a middle ground between safeguarding freedom of speech and blocking harmful misinformation effects.

6. Government Surveillance: Privacy and Civil Liberties

   The disclosure of government surveillance programs by Edward Snowden initiated debates about privacy protection and security alongside civil liberties. Ethical violations within this framework challenge the proper relationship between national security requirements and personal rights. The widespread practice of mass surveillance has created privacy concerns because it suppresses free speech while providing opportunities for governmental abuse.

In conclusion, various areas of society experience ethical violations which produce enduring and significant effects. The scenarios presented in this session demonstrate why ethical standards must guide business operations and technological innovation as well as healthcare delivery and governmental decision making. These violations lead to trust destruction and public safety threats and can trigger extensive changes to the existing systems. Society plays an essential role in creating an ethical and just future through its responses to such violations through regulatory reforms and public awareness campaigns as well as collective action.

8.4 The role of Professional Organizations in Promoting Ethical Behavior

The growing complexity of modern life demands ethical practices because they create trust while upholding integrity and responsible conduct in all industries. Professional organizations lead the ethical development of their members through standard creation and ethical conduct promotion which shapes industry ethical standards. This session examines the extensive functions professional organizations carry out to advance ethical conduct along with their value for both experts and the broader community.

1. Establishing Codes of Ethics

   Professional organizations establish ethical conduct by creating industry-specific ethical codes which they distribute to members. These codes establish thorough guidelines that define the moral standards which members must uphold. Professional organizations create ethical conduct standards through their established guidelines which provide professionals with ethical direction to maintain consistent and honest work practices.

2. Providing Educational Resources

   Professional organizations advance ethical development by providing members with educational resources combined with training programs. The resources provided by professional organizations include workshops together with seminars and online courses and publications that focus on industrial ethical dilemmas and best practices. The organizations develop professional decision-making skills through their educational resources which enables members to make ethical choices in complicated situations.

3. Fostering a Culture of Accountability

   The members of professional organizations receive essential training for ethical conduct from these organizations. Professional organizations establish ethical standards through their emphasis on ethical conduct and their presentation of ethical misconduct consequences thus enabling professionals to maintain ethical conduct towards themselves and their peers. The established accountability framework protects industry members from unethical conduct while building a favorable image for the entire sector.

4. Advocating for Ethical Policies and Legislation

   A supportive regulatory environment is essential for ethical behavior since it extends beyond individual responsibilities. Through their unified power base professional organizations work to promote policies and laws which serve both professional interests and public welfare. Professional organizations actively participate in public policy discussions while engaging with government officials to establish a regulatory framework that maintains ethical standards.

5. Mediating Ethical Dilemmas

   Professional members of complex industries must resolve numerous ethical situations which require thorough evaluation. Professional organizations act as neutral entities to assist their members who need help with ethical problems. Professionals gain help to make ethical choices through this role which provides them with perspectives and resources alongside insights that aid in ethical decision-making.

6. Peer Networking and Collaboration

   Professional organizations enable members to engage in ethical dialogue while sharing their experiences and benefiting from the knowledge of peers. These communications establish a communal atmosphere through which members exchange ethical ideas and viewpoints. The collective educational process between professionals helps them develop better comprehension of ethical matters while maintaining their dedication to ethical practices.

In conclusion, the promotion of ethical conduct by professional organizations serves as an essential element to sustain industry integrity and professional credibility in all sectors. The organizations promote responsible conduct through their establishment of ethics codes as well as educational materials and accountability measures and advocacy work and mediation services and professional networking opportunities. These organizations support ethical professional development which produces better member outcomes and benefits the entire society through ethical conduct. The essential role of professional organizations in ethical behavior guidance will continue as ethical challenges persist in the future.

8.5 Strategies for Maintaining Ethical Standards in the Technology Industry

The technology industry maintains its position as the most innovative sector because it determines the operational framework of modern society and human communication. Technological progress requires ethical standards in this industry because it continues to become more central to everyday life. The integration of technology into our daily lives depends on ethical principles to guarantee that new technologies serve social values protect individual freedoms and minimize possible risks. This session investigates various essential approaches to uphold ethical standards in the modern technology industry which operates through continuous evolution.

1. Comprehensive Ethical Frameworks

   The foundation for ethical standards in technology consists of creating and putting into practice complete ethical frameworks. These frameworks should establish essential principles alongside values and operational guidelines which both professionals and organizations need to follow. These frameworks would contain rules about data privacy alongside guidelines for algorithmic fairness and responsible AI development and societal implications of technologies. The provided ethical frameworks establish a clear direction which helps industry stakeholders make decisions and maintain appropriate conduct.

2. Rigorous Education and Training

   The technology industry depends on continuous education and training to keep its professionals knowledgeable about new ethical challenges which emerge. Organizations together with professional groups must establish training programs to teach members about data ethics and cybersecurity as well as bias mitigation and responsible innovation. Through these initiatives people acquire capabilities to detect ethical issues while they learn to make wise decisions and solve intricate ethical problems.

3. Transparent Design and Development

   The implementation of transparency stands as a vital approach for upholding ethical standards in technology. Professional teams must implement clear methods for designing and developing technologies which enable users to understand operational processes and data collection methods and usage procedures. When companies share information about their methods users develop trust and achieve better understanding to select appropriate technological involvement. Companies need to provide clear explanations about all potential threats along with restrictions that affect their products and services.

4. Ethical Impact Assessments

   Ethical impact assessments conducted before technology and feature launches help detect potential hazards along with unanticipated effects. The assessments determine how technology systems influence different stakeholders including vulnerable population segments as well as general societal groups. Professionals who identify ethical concerns before development can reduce harm while maximizing benefits by addressing these issues during the design and development process.

5. Cross-Disciplinary Collaboration

   The technology industry faces ethical challenges which demand solutions that combine insights from multiple disciplines. Working together between technologists and ethicists and legal experts and social scientists and other stakeholders produces more complete solutions to ethical problems. When experts from different disciplines work together they develop complete knowledge about technological consequences and can recognize hidden ethical issues.

6. Accountability and Reporting Mechanisms

   The implementation of strong accountability systems remains vital for preserving ethical conduct. Companies need to foster an environment which allows employees to report ethical matters without facing any negative consequences. The combination of whistleblower protections and confidential reporting channels together with defined escalation paths creates an environment that enables employees to present ethical matters. Organizations must disclose their moral obligations publicly while providing open communication about their strategies to manage ethical problems.

In conclusion, the technology industry holds significant control over human existence so it needs to maintain absolute ethical standards. The technology industry can evolve at a rapid pace while protecting societal values through the implementation of comprehensive ethical frameworks and rigorous education and training along with transparency measures and ethical impact assessments and cross-disciplinary collaboration and accountability mechanisms. The maintenance of ethical standards serves both as an ethical requirement and as a method to establish trust and drive responsible innovation which ensures technology works for human benefit.

8.6 Whistleblowing: Exposing Wrongdoing vs. Honor in Professionalism

Professional ethics depend on whistleblowers to reveal organizational misconduct because they expose hidden cases of fraud and corruption which otherwise remain hidden. People who become whistleblowers risk their personal lives and professional careers to share vital information that benefits the public. Whistleblowers drive essential reforms while creating intricate ethical and legal issues through their actions.

The ethical justification for whistleblowing exists when the following conditions apply:

1. The person reporting the matter possesses concrete proof that the organization performs unethical activities or illegal conduct.
2. The organization has dismissed or silenced all internal efforts to resolve the matter.
3. The whistleblower's actions should create positive change because evidence indicates that such outcomes are likely.
4. The disclosed information matches the necessary level required to address the identified wrongdoing.

Whistleblowing exists as an ethical practice because it allows people to fulfill their obligations to protect others from harm and support fairness and maintain professional ethics. Power needs checks through whistleblowers who protect public interests because organizations must demonstrate self-regulation and internal controls remain active.

Professional Ethics and Whistleblowing: Professional ethics represent the moral requirements and duties which belong to specific occupations. Professional ethical codes demand that members uphold the law while defending public safety and sustaining the reputation of their field. The professionals who blow the whistle usually notice that their organization performs actions which diverge from established ethical guidelines of their specific field.

Healthcare professionals must follow the ethical principles of beneficence and non-maleficence that govern their practice. When medical staff discover practices that endanger patient safety they usually decide to report the misconduct. Engineers have a duty to maintain public safety so exposing structural faults and safety lapses is considered part of their professional obligations.

People who act as whistleblowers usually face negative employer reactions that may result in job loss and career deterioration and blacklisting. Whistleblowers face various obstacles that include legal problems and social rejection as well as personal anxiety. Several countries have implemented protective laws like the Whistleblower Protection Act in the United States which establishes defense mechanisms for whistleblowers against retaliation and provides reporting systems for misconduct.

The results of whistleblower activities become significant. The process of whistleblowing results in legal prosecution of corrupt persons and organizations and leads to stronger regulations and organizational changes. The revelation of Edward Snowden and Panama Papers scandals triggered worldwide examinations of government surveillance programs and offshore banking secrecy practices.

Ethical Dilemmas: Whistleblowing presents ethical dilemmas. The main conflict emerges between the obligation to serve one's employer and the duty to serve the broader public. Professionals face confidentiality challenges since their position grants them access to sensitive information that they need to protect. Professionals who expose wrongdoing need to handle their professional responsibilities with caution when determining whether to serve their organization or protect the general public.

Digital age technologies simultaneously help and make whistleblowing more difficult. The digital age provides users with easier ways to distribute information anonymously but it also enables easier tracking of whistleblowers. People who wish to safely report wrongdoing use encryption together with secure drop services.

In conclusion, professional whistleblowers play an essential role in achieving transparency and accountability within high-stakes sectors including public health and governance and safety. Professional ethics are demonstrated through whistleblowing since individuals choose to protect public welfare and professional integrity above all else including personal and organizational interests. Society needs to establish protective frameworks that support whistleblowers while creating constructive methods to address their concerns because the implications of whistleblowing remain complex.

8.7 Case Study 1

RESEARCH ARTICLE

Exploring the ethical issues in research using digital data collection strategies with minors: A scoping review

Danica Facca[1☯]*, Maxwell J. Smith[2☯], Jacob Shelley[2☯¤a], Daniel Lizotte[3☯¤b], Lorie Donelle[4☯]

1 Faculty of Information and Media Studies, Western University, London, ON, Canada, 2 School of Health Studies, Western University, London, ON, Canada, 3 Department of Computer Science, Western University, London, ON, Canada, 4 Arthur Labatt Family School of Nursing, Western University, London, ON, Canada

☯ *These authors contributed equally to this work.*

¤a *Current address: Faculty of Law, Western University, London, ON, Canada*

¤b *Current address: Department of Epidemiology and Biostatistics, Western University, London, ON, Canada*

* *dfacca2@uwo.ca*

Abstract

While emerging digital health technologies offer researchers new avenues to collect real-time data, little is known about current ethical dimensions, considerations, and challenges that are associated with conducting digital data collection in research with minors. As such, this paper reports the findings of a scoping review which explored existing literature to canvass current ethical issues that arise when using digital data collection in research with minors. Scholarly literature was searched using electronic academic databases for articles that provided explicit ethical analysis or presented empirical research that directly addressed ethical issues related to digital data collection used in research with minors. After screening 1,156 titles and abstracts, and reviewing 73 full-text articles, 20 articles were included in this review. Themes which emerged across the reviewed literature included: consent, data handling, minors' data rights, observing behaviors that may result in risk of harm to participants or others, private versus public conceptualizations of data generated through social media, and gatekeeping. Our findings indicate a degree of uncertainty which invariably exists with regards to the ethics of research that involves minors and digital technology. The reviewed literature suggests that this uncertainty can often lead to the preclusion of minors from otherwise important lines of research inquiry. While uncertainty warrants ethical consideration, increased ethical scrutiny and restricting the conduct of such research raises its own ethical challenges. We conclude by discussing and recommending the ethical merits of co-producing ethical practice between researchers and minors as a mechanism to proceed with such research while addressing concerns around uncertainty.

Introduction

Much has been written about the ethics of conducting research with minors, due in part to the distinctive ethical issues that emerge when conducting research with this population [1, 2]. Similarly, there is an emerging body of literature about the ethics of research practices that include digital data (sometimes characterized as 'big data') collection via digital technologies (e.g., smartphones). However, the ethical dimensions, considerations, and challenges that are associated with digital data collection in research involving minors remains unclear. Notably, does research that involves the generation and/or collection of digital data among minors present unique ethical challenges? What are those challenges and how might researchers best manage

https://doi.org/10.1371/journal.pone.0237875 Used with CC 4.0 license

or mitigate the ethical dimensions and challenges within a digital technology context? This scoping review explores existing literature to understand and anticipate the ethical issues associated with collecting digitally derived research data with minors in addition to possible resolutions that can be put forward based on the reviewed literature.

Minors and Digital Data

One challenge worth noting at the outset of this review is that there is no consensus as to the definition of a minor; two broad approaches could therefore be adopted for present purposes [3, 4]. The first defines a minor as any child who has not reached the age of majority, generally thought to be under the age of 18. The problem with this definition is that the age of majority varies by jurisdiction. Two notable documents which have been extremely influential in research ethics to the extent that many countries base their regulations on their guidelines, the *Declaration of Helsinki* [5] and the *International Ethical Guidelines for Health-related Research Involving Humans* [6], discuss age of majority and minors' participation in research.

For example, within the North American context, the age of majority is 18 in some Canadian provinces like Ontario, Alberta, and New Brunswick, and 19 in other Canadian provinces like British Columbia, Nova Scotia, and the territory of Nunavut. Similar to Canada, in the United States, the age of majority also varies by state. In some U.S. states like Colorado, Idaho, and Minnesota the age of majority is 18, in others like Alabama and Nebraska it is 19, and in Mississippi it is 21. Within a European context, the age of majority in all EU Member States is 18 except for Scotland where it is 16 [7]. Additionally, the age of majority may depend on context. For example, in the Canadian province of Ontario, the age of majority is 18, but Ontario legislation dictates that one has to be 19 to purchase alcohol, suggesting that, in the context of alcohol purchases, an 18-year-old is a "minor". Further, in some EU Member States, a minor will gain full legal capacity if they are married or become pregnant before reaching the age of majority [7], suggesting that, in the context of a marital contract or pregnancy, a minor can become an adult with full legal capacity even if they are under 18.

The second approach to defining a minor is based on capacity. This is consistent with established ethical guidelines which build on the *Declaration of Helsinki* [5] and the *International Ethical Guidelines for Health-related Research Involving Humans* [6] such as the *Tri-Council Policy Statement*: *Ethical Conduct for Research Involving Humans* [8]. According to the *Tri-Council Policy Statement*: *Ethical Conduct for Research Involving Humans*, if a child is "mature sufficiently to decide on their own behalf (subject to legal requirements), the researcher must seek the children's autonomous consent in order for their participation to continue" [8]. For our purposes, it is not necessary that we resolve this issue, as our interest is in how others have addressed the issue of digital data collection in research with minors. As identified in Table 8.1 below, we adopted a broad approach, using keywords in our search intended as over-inclusive.

Table 8.1 Keywords for the academic literature search.

Ethical issues	Digital Data Collection	Research	Minors
ethic*	digital*	research*	kid*
moral*	electronic*		minor*
	online		adolescen*
	biosensor		teen*
	drone		millennial*
	device		child*
	"ocial media"		infant*

https://doi.org/10.1371/journal.pone.0237875.t001

A second challenge worth noting is specifying what exactly is meant by "digital data". For our purposes, while there may be overlap with what is otherwise termed 'big data'—defined by the volume, variety, complexity, speed and value of the data—we broadly define digital data as electronic data or information however collected [9]. As Lupton notes: "People's interactions online, their use of mobile and wearable devices and other 'smart' objects and their movements in sensor-embedded spaces all generate multiple and constant flows of digital data, often about intensely personal actions and preferences, social relationships and bodily functions and movements" [10]. It is in this data that our review is primarily interested.

Methods

A multi-disciplinary research team was established to undertake this scoping review, with expertise in computer science, digital health, ethics, law, and public health. The scoping review was conducted according to Arksey and O'Malley's [11] five stage framework which included: 1) identifying the research question/s; 2) identifying relevant studies; 3) selecting relevant studies; 4) charting the data; 5) aggregating, summarizing, and reporting the results. The research team consulted a research librarian who refined the initial search strategy and recommended databases to search given the context, subject, and population of interest.

Literature Search

To address the research question, scholarly literature was searched using four categories of keywords: *ethical issues, digital data collection, research, and minors* (Table 8.1). While initially considered, the search term 'internet' was purposefully excluded given the high volume of irrelevant articles captured in the search; the term 'internet' is so ubiquitous that it was not helpful to discriminate articles that addressed the purpose of the scoping review. Boolean search operators "AND" and "OR" were used to combine keywords between and within categories. Searches were conducted in PubMed and Scopus databases.

Articles were eligible for inclusion if they: were written in English; discussed the context of interest (i.e., research), subject of interest (i.e., digital data collection), population of interest (i.e., minors), and provided an explicit ethical analysis or presented empirical research that directly addressed ethical issues arising from the topic of interest (i.e., using digital data collection in research with minors). Upon completion of the search, duplicate articles were removed, and search results were screened by title and abstract for eligibility. Screening of abstracts was undertaken by a primary reviewer and confirmed by a second reviewer. Articles which passed the abstract screening were then retrieved in full-text and further screened for eligibility by a secondary reviewer and confirmed by two additional reviewers. Articles were summarized, and characteristics were charted, including: author(s), year of publication, database source, and ethical considerations. Reference lists of full-text articles were searched and eligible articles charted. Articles were thematically coded through multiple iterative discussions among three reviewers. A narrative account of key thematic findings is presented according to each identified theme.

Search Strategy

The search strategy and results are summarized (Fig 8.1). Searching academic databases identified 1,170 abstracts, with 1,156 remaining following the removal of duplicates. Preliminary screening of abstracts identified 73 articles for full-text screening, of which 20 articles met the inclusion criteria. Hand searching reference lists did not identify any additional relevant articles, resulting in a total of 20 articles for inclusion in the scoping review.

https://doi.org/10.1371/journal.pone.0237875

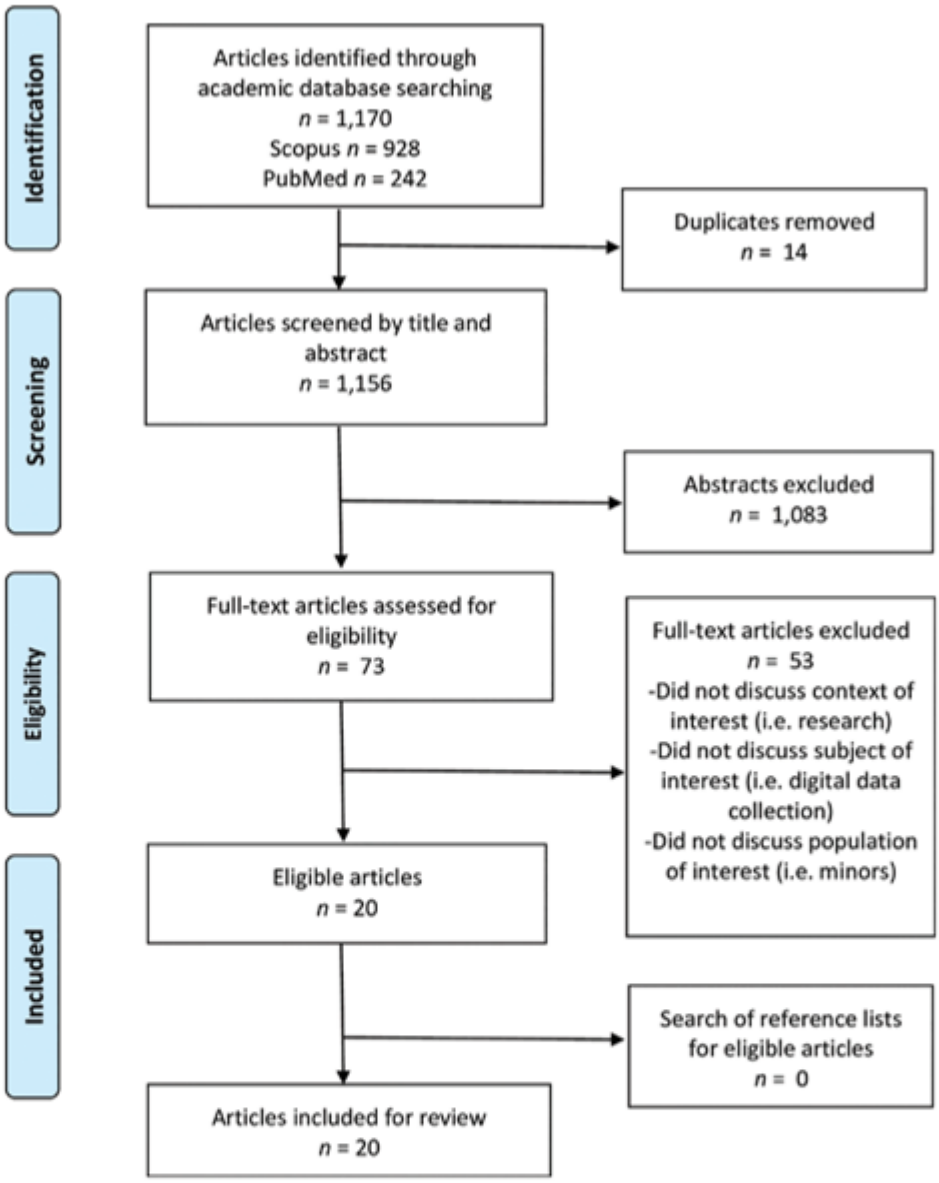


Fig 8.1 Search strategy and results.

https://doi.org/10.1371/journal.pone.0237875.g001

Scholarly Literature Characteristics

The majority of articles included in the review were published after 2014 ($n = 14$, 70%) and originated from the USA ($n = 6$, 30%), Australia ($n = 6$, 30%), and the UK ($n = 3$, 15%). More than half were original research ($n = 11$, 55%), including five qualitative studies (5/11, 45%). Overall, half of the articles ($n = 10$, 50%) focused on general ethical concerns with web- or internet-based digital data collection methods. Five articles ($n = 5$, 25%) focused specifically on the ethical concerns of using social media as a means of digital data collection. In general, social media ($n = 9$, 45%) was the most popular means of digital data collection. Facebook ($n = 8/9$, 88%) was the most popular platform identified, followed by Myspace ($n = 3/9$, 33%),

YouTube (n = 2/9, 22%), Skype (n = 1/9, 11%), and Twitter (n = 1/9, 11%). Touch-screen technology was the second most popular means of digital data collection identified (n = 6, 30%) and included smartphones (n = 5/6, 83%) and tablets (n = 3/6, 50%). The remainder of digital data collection means included online forums (n = 4, 20%), computers (n = 2, 10%), internet-based surveys (n = 2, 10%), and an electronic health record system (n = 1, 5%). Of the participant samples described in the original research articles (n = 5), the majority age range of minors was 14 years of age and under (n = 3). One study focused on minors between 14 and 17 years of age, while another study focused on minors between 7 and 17 years of age (Table 8.2).

Table 8.2 Description of scoping review scholarly articles.

First author (Year)	Geographical location	Type of Scholarly Article	Research focus	Minor age range	Means of digital data collection
Almeida (2017)	Portugal	Original research (qualitative)	Minors' use of digital devices at home and school	9–14	Desktop computer
Cleminson (2015)	United States	Report	Use of social media in research with pediatric populations	n/a	Social media (Facebook, YouTube, Skype)
Cowie (2017)	New Zealand	Original research	Ethical framework for conducting research with minors and ICTs	3.5–14	iPads, smartphone
Denissen (2010)	Germany, Sweden	Review	Psychosocial consequences of minors' internet use	n/a	Internet-based questionnaires
Donkin (2016)	Australia	Original research (qualitative)	Minors' use of virtual worlds/ games (e.g., Minecraft)	5–12	Computer, laptop, iPad
Eagleson (2017)	Canada	Original research (intervention)	Impact of mobile lifestyle intervention in overweight and obese minors with congenital heart disease	7–17	Smartphone
Fry (2014)	Australia	Original research (qualitative)	Key informant views on the ethical concerns of Australia's new PCHER system around minors	n/a	Personally controlled electronic health record (PCEHR) system

(*Continued*)

https://doi.org/10.1371/journal.pone.0237875

Table 8.2 Description of scoping review scholarly articles. (*Continued*)

First author (Year)	Geographical location	Type of Scholarly Article	Research focus	Minor age range	Means of digital data collection
Germain (2018)	United Kingdom	Review	Uses, values, benefits, and ethical concerns of online research methods with minors and adults	13–18	Online forums, social media (Facebook, YouTube)
Henderson (2012)	United Kingdom	Case study	Ethical issues in pediatric pain research with minors	n/a	Online message board, internet-based therapy intervention
Henderson (2013)	Australia	Review	Researchers' strategies used to handle ethical issues with research involving minors and social media	0–18	Social media
Hokke (2018)	Australia	Review	Ethical issues in engaging parents and minors in research online	0–18	Email, websites, online forums, social media (Facebook, Myspace)
Kia-Keating (2017)	United States	Original research (qualitative)	Ethical considerations for social media use in community-based participatory research with minors	n/a	Smartphones, social media
Lunnay (2015)	Australia	Original research (qualitative)	Social influences of drinking alcohol among minors	14–17	Social media (Facebook)
Martin-Ruiz (2018)	Spain	Methodologies	Ethical implications of minors' data protection rights with 'smart toys'	2–3	Smart toys (with accelerometers, pressure sensors, position sensors, etc.)
Moreno (2008)	United States	Original research	Ethical concerns when using social media for research purposes	n/a	Social media (Myspace, Facebook)

https://doi.org/10.1371/journal.pone.0237875

First author (Year)	Geographical location	Type of Scholarly Article	Research focus	Minor age range	Means of digital data collection
Parsons (2015)	United Kingdom	Original research	Potential of digital technologies to support minors' rights and decision-making in research participation	n/a	Touch screen technology (tablet, smartphone)
Schwab-Reese (2018)	United States	Review	Use of social media and internet-based technologies for maltreatment research	n/a	Online surveys, online forums, social media (Twitter, Facebook)
Shapiro (2013)	United States	Report	Ethical and regulatory challenges of researching online educational games	n/a	Social media (Facebook)
Spriggs (2009)	Australia	Original research	Ethical issues of minors' consent in cyber research	n/a	Social media (Myspace)
Standlee (2017)	United States	Methodologies	Ethical concerns of conducting digital ethnography with minors	n/a	Smartphones, social media (Facebook), blogs, online games

https://doi.org/10.1371/journal.pone.0237875.t002

Results

The thematic analysis generated 6 themes that addressed ethical considerations regarding digital data collection in research with minors which included: consent, data handling, minors' data rights, observing behaviors that may result in risk of harm to participants or others, private versus public conceptualizations of data generated with social media, and gatekeeping. Although these themes can generally be applied to any research that leverages digital data collection, two out of the six themes, minors' data rights and gatekeeping, warrant greater attention as they uniquely apply to conducting research when the population of interest is minors. The following sections detail the ethical issues raised according to each theme. For purposes of this review, the term *minors* is used interchangeably with any age descriptor identified in the reviewed literature (i.e., child/ren, kid/s, teen/s, adolescent/s, infant/s).

Consent

Consent is an ethical requirement when conducting research with human participants. General principles of consent under policies that inform research practices like the *Tri-Council Policy Statement*: *Ethical Conduct for Research Involving Humans* [8], mandate that an individual's consent must be given voluntarily and can be withdrawn at any time along with their data or biological materials upon request should they choose to withdraw their participation. To make such a decision, policies like the *Tri-Council Policy Statement*: *Ethical Conduct for Research Involving Humans* recommend that consent be determined by an individual's

decision-making capacity, that is, their ability "to understand relevant information presented (e.g., purpose of the research, foreseeable risks, and potential benefits), and to appreciate the potential consequences of any decision they make based upon this information" [8], rather than relying on age.

Autonomy, or the capacity needed to appreciate and understand the relevant information presented about a study in order to make an informed voluntary decision [8], is a necessary condition for consent. When it comes to involving minors in research, consent can be achieved in one of two ways: the researcher/s may obtain consent from the minor if it is determined that they have the capacity to make such an autonomous decision. Alternatively, the researcher/s may obtain consent from an authorized third party, like a parent or guardian, if it is determined that the minor does not have the capacity to make such an autonomous decision. If consent is obtained by an authorized third party, then assent to participate in the proposed research is sought from the minor. Consent by an authorized third party and assent are both needed for participation in research where it has been determined that a minor lacks autonomous decision-making capacity [8].

Numerous articles included in this review highlighted consent and assent as an ethical imperative when using digital data collection in research with minors [12–18]. Given the various geographical contexts of the research literature, the legal and ethical guidelines which determined the authors' approaches to using digital data collection with minors varied respectively. Parsons [17] drew attention to digital data collection using digital technologies and the implications of this for participant recruitment and obtaining participant consent. Parsons [17] argued that digital technologies have the potential to support prospective participants' autonomy, engagement, and decision making by: improving the accessibility of the research information; increasing motivation to take part in the study; and enhancing competency to inform decision-making. The author [17] further mentioned that digital technologies could be leveraged to make participant recruitment more inclusive, especially for minors with physical or learning disabilities. For example, using a laptop to present written text to prospective participants could give researchers the ability to tailor font size and color, background color, symbols, and computer-generated speech which can be paused, played, or slowed down, according to a minor's needs [17]. Touch interfaces, like a tablet, were also a way to enhance recruitment and consent possibilities as they tended to be intuitive to users and did "not add unnecessary complexity to the learning process" [17].

Cowie and Khoo [12] approached issues of consent when using digital data collection in research with minors whereby minors are recognized as social actors and experts in their own lives. To this end, they informed prospective minor participants of the research aims and processes using multiple strategies to support minors' decision to consent or dissent to participation in the research. They sent home a general newsletter about the study; created an interactive website about the study and research team for minors and their parents to view and discuss at home; and asked the minors' teachers to explain the research aims in school [12]. Cowie and Khoo emphasized that once consent or assent is obtained from a minor participant in any study, researchers should treat it as provisional, that is to say, "ongoing and dependent on researcher-researched and inter-participant relationships…built upon sensitivity, reciprocal trust and collaboration" [12]. Treating consent as provisional and ongoing necessitates that researchers assess participants' voluntariness at each point of contact and remind them of their right to withdraw their participation at any time [12].

Researchers [12, 17] also advocated for the use of digital technologies to facilitate the consent process as they offer "possibilities for multimodal/multimedia communication that improves the accessibility of research information beyond the inclusion of different fonts, formatting styles and images" [12]. For example, the use of digital video clips to recruit, inform, and debrief minor participants in an interactive manner engaged participants more effectively than a paper printout [12]. When considering the integral role and ethical imperative of obtaining consent from minors, Cowie and Khoo emphasized that the "onus is on researchers to use appropriate methods to achieve…consent in a way that scaffolds children's understanding and encourages and maintains their voluntary and positive participation" [12].

https://doi.org/10.1371/journal.pone.0237875

One scoping review conducted by Hokke et al. [19] which explored the ethical issues of digital data collection in research with minors (i.e., using the internet to recruit prospective participants for family and child research) also discussed the challenges of obtaining consent and assent. In their review of research, they concluded that minor consent and parental consent was more complex and ethically challenging when facilitated through online means, rather than face to face interaction [19]. Many of the studies reviewed by the authors [19] noted that obtaining consent through online means posed a risk that minors would fraudulently complete their parents' online consent form. To circumvent this risk, they [19] contacted prospective participants online, and then obtained verbal consent over the phone to assess parents' and minors' understanding of the research aims, procedures, and risks using back-questioning techniques.

Depending on the topic of interest, Hokke et al. [19] noted that waiving parental consent might in fact be a methodological and ethical necessity. For example, two US studies waived parental consent out of ethical necessity [20, 21] to collect digital data from gay and bisexual minors as they were considered to be at risk if they had to disclose their sexual identity to their parents as part of the consent process. Further, where the law required parental consent for a minor's participation in research, researchers identified this as a possible deterrent for minors' participation as some minors may be reluctant to ask their parents for permission [19].

Data Handling

One ethical imperative about the use of digital data collection with minors that was consistent throughout the reviewed literature was the importance of safeguarding data [8]. Safeguarding practices mentioned by researchers in the current review included data encryption, storage location, and secure server technology [14, 15, 22–24].

Some researchers [25] stressed that the various ethical issues at stake when handling digitally derived data depended on whether the data was actively, or passively, collected. For Schwab-Reese et al. [25], active data collection methods closely aligned with traditional data collection methods as they consist of direct interaction with research participants, even if the interaction is facilitated through electronic means [25]. Active data collection methods required participants to actively engage in the data collection process; in direct conversation with the researcher, or in responding to survey questions. For example, a researcher may ask a series of questions in real-time through a social media app to which a participant can give an immediate response. Alternatively, passive data collection methods aligned with secondary data analysis as they do not require direct interaction with participants, but rather, aggregate and analyze large sets of existing data [25]. For example, a researcher may monitor the number of hours a participant spends on a certain app, like Twitter, and therefore does not need to be in direct contact with the participant to collect that data in real-time.

Since a researcher would have direct contact with participants when using active data collection methods, informed consent or assent processes follow suit as standard practice [25]; the participant would be made aware that whatever data they produced in that given research context would be given over to the researcher for analysis. Examples of active digital data collection methods include crowdsourcing (i.e., most commonly done through websites where an open call for participation is issued for users to complete a short survey), or, online recruitment (i.e., most commonly done through websites or social media). Using active digital data collection methods in research with minors may not pose any additive ethical issues beyond those mentioned earlier when considering consent/assent processes as well as the degree to which the data are safeguarded through encryption, server technology, and storage location.

However, passive data collection methods proved to be more ethically challenging for researchers [25] when considering users' privacy expectations about their digital data. Since passive data collection methods, according to Schwab-Reese et al. [25], are predicated on analyzing existing data, this means researchers may not have to necessarily seek out consent or assent from users which could be considered ethically problematic

https://doi.org/10.1371/journal.pone.0237875

(e.g., does the user consider their digital data on a social media platform, like Twitter, private or public information?). Methods of passive digital data collection include internet search queries (e.g., Google Search Trends, which regularly updates a database of aggregated internet queries), forum postings (e.g., as found on a website like Reddit, or social media platform like Facebook), or social media user activity (e.g., analyzing Twitter users' tweets, likes, retweets, use of certain hashtags, etc.).

To address the ethical queries related to the use of passive digital data collection methods with minors, researchers consulted guidelines from the Association of Internet Researchers [26] that recommended consideration of the following: what constitutes 'engaging with a human subject'; definitions of private and public space; and finally, tensions between regulatory and context specific ethical decision making [27].

Online forums and social media platforms offer vast research opportunities for publicly available and passively collected data. On the surface, research processes required for ethical research 'engagement with a human subject' would not seemingly apply to the collection of publicly available online data. However, Schwab-Reese et al. noted that while passive data collection studies typically dealt with "aggregated or de-identified data… recent research suggests de-identified datasets often contain sufficient personal information to potentially identify individuals" [25]. Current recommendations for the ethical practice of online and publicly available passive data collection may exempt researchers from the customary oversight of an institutional ethics review board or regulatory body. Instead, researchers are recommending the establishment of "an external advisory committee to critically process the potential harms, vulnerabilities, benefits, and so forth, even if the research does not directly engage with human subjects and thus, does not explicitly require intuitional ethics review" [25]. They suggest this may be best practice moving forward in order to deal with research that digitally collects "person-based data without direct human contact" [25].

Ethical research practices pertaining to defining public and private space with respect to digital data collection methods should "carefully consider the expectations of the individuals who are creating the data" [25]. Different from active digital data collection methods, where prospective participants would be informed as to how their digital data was to be used for research purposes, this may not necessarily be the case with passive digital data collection methods. For example, "the creator of a public blog or social media profile which is viewable by anyone may have a lower expectation of privacy than in a private forum that requires log-in by members" [25]. Therefore, if researchers seek to passively collect digital data from minor participants, a combined approach of both active and passive data collection methods may be appropriate to understand the privacy expectations of prospective participants (active) before using their preexisting data or monitoring their behavior (passive).

Lastly, Schwab-Reese et al. acknowledged that "regulations are often intended to encourage ethical research and practice, but when applied universally without consideration, regulations may inadvertently restrict important, necessary research" [25]. With the increase of digital derived data collection, Schwab-Reese et.al. concluded that "as technology-based research moves forward, it will be important to establish firm ethical boundaries for some clearly defined issues, while encouraging flexibility and situation-based ethical decision-making for ethically grey areas" [25]. Ethical decision making on a case-by-case basis appears to be the way forward when considering the multitude of grey areas that arise when involving minors in research that uses digital data collection, whether active or passive in nature.

Benefits of using digital data collection were reported as an efficient, low cost, environmentally friendly, and convenient way to engage with a diverse range of participants across geographical boundaries [28]. However, challenges of digital data collection methods included technical aspects like data storage (servers) and security (encryption) measures, particularly as they related to the collection of personal health information [23, 24, 28]. Two studies mentioned a law or regulation with respect to the handling of personal health information collected through digital means [23, 24]. One study provided an extensive

https://doi.org/10.1371/journal.pone.0237875

discussion related to necessary actions taken to comply with Ontario's Personal Health Information Protection Act [29] which safeguards the collection and use of personal health information [23]. Martin-Ruiz et al. [24] reported on the use of the Data Protection Impact Assessment [30], a European regulation which assesses a study's potential risks to individual privacy [30]. In their report, the researchers addressed the six protection goals regarding the collection and use of participants' personal data [24] which included elements such as data availability, integrity, confidentiality, unlinkability, transparency, and intervenability [30]. As exemplified in both of these studies, addressing the ethical issues associated with handling digital data took both regulatory bodies and context-specific (personal health information) ethical decision making into consideration.

Minors' data rights. The extent to which minors are included in discussions regarding the dissemination of their digital data for research purposes, or the sharing of their digital data with adult actors such as their parents or guardians generated through research, was frequently flagged as an ethical issue by many [12, 13, 16, 18, 22, 24, 31]. The contemporary issue of minors' (digital) data rights is part of a much larger rights movement influenced by the establishment of the United Nations Convention on the Rights of the Child in 1989 [13, 22]. Historically, minors have been framed as a marginal group, with little, if any, opportunity to weigh in on matters (like research) that affect their social worlds [22]. The establishment of the convention signaled "the deconstruction of the protectionist paradigm of childhood [and] assigned [minors] the right to be co-producers of science and involved in all stages of research on and with them" [22] and consequently led to the emergence of childhood studies as a discipline. Opportunities for minors to be consulted and included in decision making processes is recognized by childhood studies' scholars as a way to advance minors' agency and autonomy within the research landscape [13].

Another ethical consideration in relation to minors' data rights included privacy expectations of their parents/guardians [18]. Cultural and legal practices tend to normalize parent/guardian access to minors' data held within government, healthcare or educational institutions; as such, parents/guardians may expect to have access to minors' data collected for research purposes [18]. Drawing attention to minors' data rights acknowledges their right to privacy over their personal data, both in an online and offline context, along with the need to obtain their consent in order to share their data with third party adult actors like their parents/guardians [32].

Observing Behaviors that may Result in Risk of Harm to Participants or Others

Three studies mentioned the ethical challenges of reporting risky behaviors, or potential behaviors that could result in risk, when using digital data collection in research with minors [15, 33, 34]. In general, conducting research with human participants "includes the risk of observing or being informed about behavior that is illegal, amoral, immoral, or otherwise illicit" [33], to which researchers have to determine a course of action. Using digital data collection methods could potentially involve collecting information about minors' online activity, for instance on social media, which may fall outside of a research study's objectives but nonetheless be considered as illegal, amoral, immoral, or otherwise illicit. Under ethical guidelines like the *Tri-Council Policy Statement*: *Ethical Conduct for Research Involving Humans* such instances are known as material incidental findings as "they are reasonably determined to have significant welfare implications for the participant or prospective participant" [8].

Researchers [15, 33] highlighted the ongoing tension between existing laws and research ethics when conducting social media related research with minors. One study recognized that observing and responding to behaviors through social media which may harm participants or others poses complex challenges for researchers as it may be hard to decipher the parties involved (i.e., perpetrators, victims), the nature of the activity, or the proper contact to report the activity to [33]. For example, if a minor participant disclosed to a researcher that they illegally downloaded music through a social media platform, the researcher would need

to decide if they would notify the participant and report the illegal activity, even if it meant jeopardizing their trust [33]. Similarly, researchers utilized the social media platform Facebook to contact under-aged women in Australia to explore the social influences of drinking alcohol [15]. Ethical challenges emerged for researchers regarding their responsibility to report incidents of illegal, amoral, immoral, or otherwise illicit behavior (i.e., photographs of inappropriate behaviors such as nudity, illegal activity) captured via online data collection against the promise of information confidentiality. Potentially reportable data was captured despite researchers' caution to participants that "all information provided to the researcher would be treated in the strictest confidence [unless] the researcher was legally obliged to disclose information related to illegal activity if requested by relevant authorities" [15].

The ethical tension for researchers arises 'in-between' the requirement to disclose reportable incidents while ensuring an ethical duty of participant confidentiality whereby information obtained from a participant will remain confidential within and beyond a study [8]. While a researcher's promise of confidentiality is central to developing and maintaining trusting relationships with participants, it "must, at times, be balanced against competing ethical considerations or legal or professional requirements that call for disclosure of information obtained or created in a research context" [8]. Depending on the topic and objective of a study, "researchers are expected to be aware of ethical codes…or laws…that may require disclosure of information they obtain in a research context" [8]. Breaching confidentiality by reporting confidential information risks losing the trust built between researcher and participant, but in some cases, it may be necessary in order to serve the greater good by protecting "the health, life or safety of a participant or a third party, a community, or the general population" [8].

Digitally collected data highlights unique tensions for researchers to maintain the integrity of the research purpose, to honor the confidentiality of participant information, and to disclose reportable events. For example, researchers [34] had to weigh the costs of constraining participants' photographic activity against their study's goal of equity and empowerment. They concluded that it was important "to allow youth to depict the reality of the challenges that they or their community faced without necessary constraint" [34]. Navigating the risk of harm to participants and their community proved to be complex as the authors realized minors often took photos on their own mobile phones rather than the digital cameras provided to them and uploaded the photos to their personal social media accounts to be shared on those platforms [34].

Private Versus Public Conceptualizations of Social Media

A large part of the reviewed literature discussed the ongoing debate of whether social media is considered private or public domain when using digital data collection in research with minors [13, 15, 19, 25, 33–37]. Many articles drew attention to the shifting nature of privacy within digital spaces since "any information posted on the Internet [technically] enters a virtual public space" [15] even if it is posted with the intention of remaining private to certain audiences. Understandings of privacy depend on the context in which it is invoked as well as a user's expectations within that context. Since there is not a clear distinction between what is considered public and private domain within the internet, and social media more generally, maintaining user privacy and limiting potential harm surfaced as key ethical challenges [15, 33, 35, 36].

When using social media for research purposes, one study recommended researchers educate and ensure that their participants understand the Terms of Service associated with the social media site being used for the research project as it may lead to instances of public disclosure [33]. Another study stated that it is the researcher's responsibility to also learn the nuances of the site's privacy settings in order to inform their participants of the ways in which their information will be handled and, to a certain extent jeopardized, since privacy settings are not always fail-safe [15]. It was also mentioned that researchers should consider the cultural norms of the participants creating the data as well as the norms surrounding the social media platform being utilized to gauge what privacy means within these specific contexts [25].

https://doi.org/10.1371/journal.pone.0237875

When considering the cultural norms of the participants involved in research with social media, two studies recognized that "digital natives" [19], in other words, minors who are born into digital environments and learn how to use digital technologies from a young age, may conceptualize privacy differently than adults. For example, one article queried whether minors disclose private information on social media "without understanding or considering the permanence or far-reaching nature of online content, and without intending for their information to be used by others" [19]. While another mentioned that minors may disclose their private information on social media out of "naivety or ignorance" [37] rather than an intended disregard for their privacy and personal information.

Gatekeeping

By far the most acknowledged ethical issue pertaining to the use of digital data collection in research with minors was gatekeeping by a parent/caregiver [12, 17, 19, 37], relevant stakeholders like medical professionals or educators [14, 18, 31] and research ethics boards [23, 25, 33, 36, 38]. While the inherent power imbalance between minors and adult researchers within the research context necessitates the need for gatekeeping by research ethics boards [12, 14], it is important to note that "the power [parents have] as gatekeepers in the processes of…research participation…should not be under-estimated" [17]. Parents/guardians ultimately provide access to minor participants, and in some cases, consent for their participation.

Conducting research with minors, irrespective of data collection method, must account for the triadic nature of the prospective research relationship which consists of the researcher, minor, and gatekeeper. Ultimately, it is the gatekeeper who grants a researcher access to a minor's world [22]. Depending on the gatekeeper's relationship to the minor, whether they are a parent/guardian, educator, or REB member, a minor's prospective participation in a study will hinge on whether it is found to be in their best interest by these gatekeepers, and whether these gatekeepers are willing to cooperate with a researcher to make the study happen. When it comes to making decisions within the research context between adults, researchers, and minors, one authors suggests that research "decision-making needs to be understood as part of a discussion or dialogue between young people, parents/caregivers and the researchers" [17].

Discussion

The reviewed literature identified numerous ethical issues related to conducting digital data collection in research with minors which included: consent, data handling, minors' data rights, observing behaviors that may result in risk of harm to participants or others, private versus public conceptualizations of social media, and gatekeeping. While these ethical issues are pertinent to any discussion of using digital data collection in research, whether they are specifically unique when conducting such research with minors as the population of interest requires deeper consideration.

Although consent arose as a common theme among many of the studies, it is not an ethical concern unique to research with minors. Consent is, at the very least, a minimum requirement for research participation. Even so, seeking assent is not a unique ethical practice for research with minors as its tenets of accountability, reciprocal trust, and consultation should be considered an ethical necessity for all populations involved in research. When using digital data collection in research with minors, specifically within the context of social media, researchers [33] highlighted the importance of maintaining an open dialogue, or as they called it, a *dialogic approach*, throughout the research process. Following a dialogic approach with minors when using social media as an avenue for digital data collection allows researchers to assess minors' feelings about their posts to social media and understandings of how published research will "transform…their private information and interactions into public data" [33]. Maintaining a continual dialogue with minor participants reaffirms the notion of provisional consent and that minors' participation and data can be withdrawn at any time upon request.

https://doi.org/10.1371/journal.pone.0237875

While data handling and ownership are not unique ethical challenges when conducting research, they do warrant greater attention in a digital context. Since digital data encryption, secure storage location, and access are all ethical challenges which equally apply to research with minor and adult populations, the main ethical concern with data handling in this review was the extent to which the digital environment creates concomitant opportunities for data breaches. As noted by some researchers, "the very nature of the internet introduces security and privacy issues, including potential privacy breaches through hacking and data corruption during transfer" [23]. Given that security risks within a digital context are ever-present, a top priority for researchers using digital data collection methods will be ensuring, as is possible, that no manipulation, in other words hacking, happens during the transmission, encryption, and storage of participants' digital data [24].

Observing behaviors which assume risk, or may result in risk to participants or others, is a conventional part of conducting research with human participants. While the reviewed literature acknowledged the ethical complexity of navigating such reporting in relation to participants' social media data, deciphering what the activity is, what parties are involved, and whether there is an authority that the activity is required to be reported to, are ethical issues which can arise in any study with any population. As mentioned by some researchers, employing digital data collection methods warrants "an ongoing dialogue amongst researchers and ethics committees and between researchers and participants around…dilemmas, as well as processes to resolve them" [33]. They noted that the "constant evolution of technologies (such as social media and search capabilities) and social practices to which they are put…means that researchers and ethics committees are not necessarily equipped to understand the consequence or implications of their research practice" [33]. Therefore, within each individual research context that employs social media as a means to collect participants' digital data, it is important for researchers to understand the social media practices of their population and approach dilemmas by applying all of the perspectives from those involved [33].

Varied conceptualizations of whether content shared through social media falls within the public or private domain is an important ethical issue when examining digital data collection methods, however, it is not unique to conducting research with minors. Although some articles claimed that digital natives may be naive or ignorant of social media privacy settings and the implications of sharing private information within digital spaces, adult users face similar struggles. With the rapid development, uptake, and variety of social media platforms across the globe, nuances of handling one's personal information online affects users of all ages, backgrounds, and digital literacy levels; this is especially so given the lengthy and complex jargon of Terms of Service agreements [33]. Researchers will not necessarily face more struggles when conducting research with minors using digital data collection methods, such as social media, since users of all ages may not understand the intricacies and shifting nature of online contexts to the same degree [39].

Managing Minors' Data Rights

Although minors' data rights seems to point toward a unique ethical issue when conducting research with minors using digital data collection, many of the issues raised can be extended to other populations. For instance, weighing in on the treatment and dissemination of personal data for research purposes can advance the agency and autonomy of any aged participant. This is especially so given that the figure of the child tends to be positioned as "exceptional…rather than part of the wider frame of rights and the digital" [32]. While there may not be unique ethical concerns, in that other populations also have their best interests dictated by others such as the old, poor, or disabled, one might argue that where minors might differ is in their potential to be involved with research decision-making [32].

Addressing Gatekeeping in Research Involving Minors and Digital Technology

Institutional review boards, or research ethics boards (REBs), dictate the scope of any research project for they are most concerned with risk mitigation [18]. Although REBs seek to reduce potential harm to participants while preserving the potential benefits of a research endeavor, they "can be a problematic gatekeeper for

researchers, especially for those who are seeking to conduct research in new and contentious areas like online spaces" [33]. Since there is an additive degree of uncertainty that invariably exists when considering research that involves minors as well as digital technology, REBs themselves "may not even be equipped to best guide ethical practices concerning these new areas of inquiry" [33].

Gatekeeping is a unique ethical issue concerning research with minors that involves digital data collection. Considering that minors' use of digital technology in general raises ambivalence among adults, it is no surprise that minors' participation in research with digital technologies faces hypervigilant gatekeeping. As one author writes, research "decision-making needs to be understood as part of a discussion or dialogue between young people, parents/caregivers and the researchers" [17]. Perhaps the ethical issue at stake is not the population or topic of interest per se, but rather, the way in which gatekeeping may impede research efforts that can provide opportunities for minors to inform and shape understandings of our uncharted digital environments.

Limitations

This review has several limitations. First, although the search strategy intended to be inclusive with its terminology, there may be relevant articles that were not captured. For example, articles may have described the context, subject, and population of interest without using 'ethic' or 'moral' in their title or abstract. This limitation also extends to the hand searched reference lists as only titles were screened for inclusion. Second, the search strategy was limited to articles written in English so work which may have contributed to this ethical discussion that was written in another language was not included. Third, the search strategy did not focus on one type of digital data to be collected (e.g., health data, screen-time data) which, if refined, could have identified further ethical issues or gaps in the literature. However, this was intentionally left outside of the scoping review search strategy as we were primarily concerned with the ethical issues of using digital data collection in research with minors, rather than the type of data being collected through digital means.

Conclusion

As indicated at the outset of this review, our intention was to explore existing literature to understand and anticipate the ethical issues associated with collecting digitally derived research data with minors in order to forward any possible resolutions based on the reviewed literature. The reviewed literature indicated that there was no difference in ethical issues when collecting digitally derived research data with minors in comparison to other populations except for gatekeeping. Gatekeeping is a unique ethical issue when collecting digitally derived research data with minors given that it is both a necessary safety measure to ensure minors are not taken advantage of within the research context *and* a potential barrier to minors' participation since digital technology is a contentious area of research. For our purposes, this ethical conundrum begets the following question: if gatekeeping is a necessary barrier to minors' participation in research which specifically involves the collection of digitally derived research data, how do we resolve this as researchers based on the reviewed literature? Our resolution to this conundrum is to suggest that researchers co-produce ethical practice with minors.

At various points throughout the reviewed literature, it was continually suggested that any ethical issue associated with digital data collection in research with minors may be best addressed when minors are part of the research conversation and decision-making processes [12,13, 16–18, 22, 24, 31, 33]. In this sense, co-producing ethical practice with minors is the most respectable resolution to approaching ethical dilemmas in an area of research where the technology itself, along with accompanying social practices, are constantly evolving. Co-producing ethical practice in research which collects digitally derived research data from minors could be addressed by implementing a Child and Youth Advisory Committee (CYAC) [37]. CYACs seek to balance children's protection while supporting their participation in research [40]. In particular, CYACs have

been implemented in research with minors which addressed similarly contentious topics including cyber safety [41], hazardous agricultural labor [42], self-advocacy for pediatric patients with chronic illness [43], and child rights [40].

The problem when using digital data collection in research with minors is not necessarily the minors, nor the digital technology, but the uncertainty surrounding it. Conducting research with minors, along with digital technology, compounds uncertainty and increases ethical scrutiny. Uncertainty should not lead to preclusion, but rather, to co-production of ethical practice between researchers and minors. Co-production *is* risk mitigation; it is not an antidote to risk but an approach to working in tandem with minors to foster best ethical practice when using digital means to collect their data within a research context.

Supporting Information

S1 File.
(PDF)

Author Contributions

Conceptualization: Danica Facca, Maxwell J. Smith, Jacob Shelley, Daniel Lizotte, Lorie Donelle.
Data curation: Danica Facca, Maxwell J. Smith, Jacob Shelley, Daniel Lizotte, Lorie Donelle.
Formal analysis: Danica Facca, Maxwell J. Smith, Jacob Shelley, Daniel Lizotte, Lorie Donelle.
Supervision: Lorie Donelle.
Writing – original draft: Danica Facca, Maxwell J. Smith, Jacob Shelley, Daniel Lizotte, Lorie Donelle.
Writing – review & editing: Danica Facca, Maxwell J. Smith, Jacob Shelley, Daniel Lizotte, Lorie Donelle.

References

1. Bos W., Tromp K., Tibboel D., Pinxten W. Educational Paper: Ethical aspects of clinical research with minors. Eur J Pediatr. 172:859–866. https://doi.org/10.1007/s00431-012-1856-8
2. Fargas-Malet M., McSherry D., Larkin E., Robinson C. Research with children: methodological issues and innovative techniques. J Early Child Res.2010;8(2):175–192. https://doi.org/10.1177/1476718X09345412
3. Campbell T.D. The rights of the minor: As person, as child, as juvenile, as future adult. Int J Law Fam.1992;6(1):1–23. https://doi.org/10.1057/cpt.2011.43
4. Cowden M. Capacity, claims and children's rights. CPT.2012;11:362–380. https://doi.org/10.1057/cpt.2011.43
5. WMA Declaration of Helsinki–Ethical Principles for Medical Research Involving Human Subjects. World Medical Association. 2018 July. Available from: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/
6. International Ethical Guidelines for Health-related Research Involving Humans, Fourth Edition. Geneva. Council for International Organizations of Medical Sciences (CIOMS); 2016. Available from: https://cioms.ch/wp-content/uploads/2017/01/WEB-CIOMS-EthicalGuidelines.pdf

7. Mapping minimum age requirements concerning the rights of children in the EU. European Union Agency for Fundamental Rights. 2017 Nov. Available from: https://fra.europa.eu/en/publication/2017/mapping-minimum-age-requirements/age-majority

8. Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council. Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans. 2018 Dec. Catalogue No: RR4-2/2019E-PDF. ISBN: 978-0-660-29942-6.

9. Hiba Jasmin H., Ammar Hameed S., Hadisheheed S., Azizahbt Haji A. Big Data and the Five V's Characteristics. IJAECS.2015;2(1):16–23. https://doi.org/10.1007/978-3-319-06811-4_13

10. Lupton D. Feeling your data: Touch and making sense of personal digital data. New Media Soc.2017;19(10):1599–1614. https://doi.org/10.1177/1461444817717515

11. Arksey H., O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol.2005;8(1):19–32. https://doi.org/10.1080/1364557032000119616

12. Cowie B., Khoo E. Accountability through access, authenticity and advocacy when researching with young children. Int J Inclu Edu.2017;21(3):234–247. https://doi.org/10.1080/13603116.2016.1260821

13. Donkin A., Holloway D., Green L. Towards a participatory Netnography: collaborating with children in virtual worlds research. J Med Comm.2016;7:5–16. Available from: https://platformjmc.files.wordpress.com/2016/10/adonkin_netnography_fullpaper.pdf

14. Henderson E.M., Law E.F., Palermo T.M., Eccleston C. Case Study: Ethical Guidance for Pediatric e-health Research Using Examples From Pain Research With Adolescents. J Pediatr Psychol.2012;37(10):1116–1126. https://doi.org/10.1093/jpepsy/jss085 PMID:22851643

15. Lunnay B., Borlagdon J., McNaughton D., Ward P. Ethical Use of Social Media to Facilitate Qualitative Research. Qual Health Res.2015;25(1):99–109. https://doi.org/10.1177/1049732314549031 PMID:25212856

16. Moreno M.A., Fost N.C., Christakis C.D. Research Ethics in the Myspace Era. Pediatrics.2008;121(1):157–161. https://doi.org/10.1542/peds.2007-3015 PMID:18166570

17. Parsons S. The Potential of Digital Technologies for Transforming Informed Consent Practices with Children and Young People in Social Research.JoSi.2015;3(6):56–68. https://doi.org/10.17645/si.v3i6.400

18. Standlee A. Digital Ethnography and Youth Culture: Methodological Techniques and Ethical Dilemmas. Social Stud Child Youth.2017;22:325–348. https://doi.org/10.1108/S1537-466120180000022015

19. Hokke S., Hackworth N.J., Quin N., Bennetts S.K., Win H.Y., Nicholson J.M. et al. Ethical issues in using the internet to engage participants in family and child research: A scoping review. PLoS ONE.2018;13(9):1–30. https://doi.org/10.1371/journal.pone.0204572

20. Hoffman N.D., Freeman K., Swann S. Healthcare preferences of lesbian, gay, bisexual, transgender and questioning youth. J Adolesc Health.2009;45(3): 222–229. https://doi.org/10.1016/j.jadohealth.2009.01.009 PMID:19699417

21. Nielsen S., Paasonen S., Spisak S. 'Pervy role-play and such': girls' experiences of sexual messaging online. Sex Educ.2015;15(5):472–85. https://doi.org/10.1080/14681811.2015.1048852

22. De Almeida A.N., Carvalho D., Delicado A. Accessing Children's Digital Practices At Home Through Visual Methods: Innovations and Challenges. Sociol Stud Child Youth.2017;22:349–374. https://doi.org/10.1108/S1537-466120180000022016

23. Eagleson R., Altamirano-Diaz L., McInnis A., Welisch E., De Jesus S., Prapavessis H. et al. Implementation of clinical research trials using web-based and mobile devices: challenges and solutions. BMC Med Res Methodol.2017;17(43):1–8. https://doi.org/10.1186/s12874-017-0324-6 PMID:28302050

24. Martin-Ruiz M.L., Fernandez-Aller C., Portillo E., Malagon J., del Barrio C. Developing a System for Processing Health Data of Children Using Digitalized Toys: Ethical and Privacy Concerns for the Internet of Things Paradigm. Sci Eng Ethics.2018;24(4):1057–1076. https://doi.org/10.1007/s11948-017-9951-x PMID:28815460

25. Schwab-Reese L.M., Hovdestad W., Tonmyr L., Fluke J. The potential use of social media and other internet-related data and communications for child maltreatment surveillance and epidemiological research: Scoping review and recommendations. Child Abuse Negl.2018;85:187–201. https://doi.org/10.1016/j.chiabu.2018.01.014 PMID:29366596

26. Buchanan E. Readings in Virtual Research Ethics: Issues and Controversies. Information Management.2005 Spring;18(1):26–27. Available from: https://www.researchgate.net/publication/293512050_Readings_in_virtual_research_ethics_Issues_and_controversies/stats

27. Markham A., Buchanan E. Ethical decision-making and internet research: Recommendations from the AoIR ethics working committee (version 2.0).2012 Summer. Available from: http://aoir.org/reports/ethics2.pdf

28. Denissen J.J.A., Neumann L., van Zalk M. How the internet is changing the implementation of traditional research methods, people's daily lives, and the way in which developmental scientists conduct research. Int J Behav Dev.2010;34(6):564–575. https://doi.org/10.1177/0165025410383746

29. Personal Health Information Protection Act, 2004, S.O. 2004, c. 3, Sched. A. (March 25, 2020). Available from: https://www.ontario.ca/laws/statute/04p03

30. Data Protection Impact Assessment. (March 12, 2019). Available from: https://edpb.europa.eu/our-work-tools/our-documents/topic/data-protection-impact-assessment-dpia_en

31. Fry C.L., Spriggs M., Arnold M., Pearce C. Unresolved Ethical Challenges for the Australian Personally Controlled Electronic Health Record (PCEHR) System: Key Informant Interview Findings. AJOB Empir Bioeth.2014;5(4):30–36, https://doi.org/10.1080/23294515.2014.919972

32. Livingstone S., Third A. Children and young people's rights in the digital age: An emerging agenda. New Media Soc.2017;19(5):657–670. https://doi.org/10.1177/1461444816686318

33. Henderson M., Johnson N.F., Auld G. Silences of ethical practice: dilemmas for researchers using social media. Educ Res Eval.2013;19(6):546–560. https://doi.org/10.1080/13803611.2013.805656

34. Kia-Keating M., Santacrose D., Liu S. Photography and Social Media Use in Community-Based Participatory Research with Youth: Ethical Considerations. Am J Community Psychol.2017;60:375–384. https://doi.org/10.1002/ajcp.12189 PMID:28944473

35. Germain J., Harris J., Mackay S., Maxwell C. Why Should We Use Online Research Methods? Four Doctoral Health Student Perspectives. Qual Health Research.2018;28(10):1650–1657. https://doi.org/10.1177/1049732317721698 PMID:28745106

36. Shapiro R.B., Ossorio P.N. Regulation of Online Social Network Studies. Science.2013;339 (6116):144–145. https://doi.org/10.1126/science.1219025 PMID:23307724

37. Spriggs M. Consent in Cyberspace. Monash Bioeth Rev.2009;28(4):32.1–32.15. https://doi.org/10.2104jmber0932

38. Cleminson J. Child health research and social media. Arch Dis Child Educ Pract Ed.2015;100:331–332. https://doi.org/10.1136/archdischild-2014-308100 PMID:25824892

39. Obar J.A., Oeldorf-Hirsch A. The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services. Inf Comm Soc.2020;23(1):128–147. https://doi.org/10.1080/1369118X.2018.1486870

40. Collins T.M., Jamieson L., Wright L.V.H., Rizzini I., Mayhew A., Narang J., et al. Involving child and youth advisors in academic research about child participation: The Child and Youth Advisory Committees of the International and Canadian Child Rights Partnership. Child Youth Serv Rev.2020;109:1–9. Retrieved from: https://doi.org/10.1016/j.childyouth.2019.104569

41. Lester L., Cross D., Terrelinck D., Falconer S., Thomas L. Encouraging the positive use of technology through community engagement. Safer Communities.2016;15(3):134–141. https://doi.org/10.1108/SC-11-2015-0035

42. Arnold T.J., Malki A., Ibarra J., Daniel S.S., Ballard P.J., Sandberg J.C., et al. Engaging Youth Advocates in Community-Based Participatory Research on Child Farmworker Health in North Carolina. PCHP.2019;13(2):191–199. Retrieved from: https://doi.org/10.1353/cpr.2019.0019 PMID:31178454

43. Rich C., Goncalves A., Guardiani M., O'Donnell E., Strzelecki J. Teen Advisory Committee: Lessons Learned by Adolescents, Facilitators, and Hospital Staff. Pediatr Nurs.2014;40(6):289–296. Retrieved from: https://pubmed.ncbi.nlm.gov/25929124/ PMID:25929124

8.8 Case Study 2

RESEARCH ARTICLE

The relationship between hospital ethical climate and continuing education in nursing ethics

Ayaka Okumoto [1]*, Satoko Yoneyama[1,2], Chiharu Miyata[3], Ayae Kinoshita[1]

1 Department of Human Health Sciences, Graduate School of Medicine, Kyoto University, Kyoto, Japan, 2 Department of Neuropsychiatry, Kanazawa Medical University, Ishikawa, Japan, 3 Course of nursing science, Mie University Graduate School of Medicine, Mie, Japan

* *okumoto.ayaka.l23@kyoto-u.jp*

Abstract

Background

In recent years, there has been a growing interest in the importance of creating a healthy ethical climate. Although relationship with various factors and the ethical climate have been reported, understanding of the relationship between ethical education and ethical climate is limited.

Aim

This study aims to investigate the relationship between ethical climate, personal characteristics, and continuing education for ethics.

Methods

This study conducted a quantitative cross-sectional survey of 605 nurses in 3 teaching hospitals in Japan. Multiple-regression analysis was used to assess the relationship between ethical climate and demographic characteristics and continuing education. Further mean of ethical climate scores were compared between received continuing education and did not, using analysis of covariance adjusted for demographic variables.

Findings

The ethical climate showed significant association with hospital, gender, specialty of the unit, experience of ethics education, in-service ethical training, and workshops/ academic conferences on nursing ethics. In multiple-regression analysis, attending in-service ethical training increased the mean of ethical climate score ($p = 0.031$) and workshops/ academic conferences decreased the mean score ($p = 0.028$). Adjusted-mean of ethical climate score of nurses who had in-service training was significantly higher than those who had not ($p = 0.038$), whereas adjusted-mean of it of nurses who had attended workshops/ academic conferences was significant lower ($p = 0.033$).

Discussion

In-service training on ethics was associated with the positive ethical climate. Hospital should enhance ethical education.

https://doi.org/10.1371/journal.pone.0269034 Used with CC 4.0 license

Conclusion

Ethical climate related to the nurses' personal characteristics and continuing education. We propose that organizational support for ethical education may be effective in raising the ethical climate of the workplace.

Introduction

In recent years, there has been a growing interest in the importance of creating a healthy and ethical work environment in clinical practice [1]. The ethical climate is considered an essential component of the organizational climate and is defined as shared perceptions of what is ethically correct behavior and ways to handle ethical issues [2]. Ethical climate refers to an environment in the organization that an individual perceives to influence and set the standard for employee behavior [3].

To measure ethical climate objectively, researchers have developed various theories and measurements. In this study, the scale developed by Olson [4], the Hospital Ethical Climate Survey (HECS), was used to measure the ethical climate of hospitals. Using instruments, it has been argued that the ethical climate may influence nurses' well-being in the workplace and the quality of patient care provision, making it a significant factor in the nursing environment [5]. Various studies show that an ethical climate may be associated with or affect moral distress [6–9], job satisfaction [10, 11] turnover intentions and medical error experience [12], satisfaction with the quality of care and collaboration [7], ethical leadership [11], and moral sensitivity [13, 14]. Additionally, research is being conducted to clarify the relationship between ethical climate and demographic factors such as age, gender, nurses' education degree, material status, years of experience, employment status, job position, education degree, training related to ethics, type of work (daytime or shift), weekly work time, the average number of patients cared for, working conditions, etc. [11, 12, 15–18].

The ethical climate is the pervasive moral atmosphere of a social system, characterized by the shared perceptions of right and wrong, as well as common assumptions about how moral concerns should be addressed [19]. Therefore, investigating the ethical climate of hospitals contributes to enhancing the work environment and quality of care in the context of nursing [18]. The ethical climate is the shared perception formed in the nurses' workplace. According to our previous study, nurses' perceptions of bed alarm system and physical restraint were related to the hospital they belong to, years of experience and education [20]. Ethics education is a method for preventing or addressing moral distress [21] and has a positive influence on moral confidence, moral action [22], ethical decision making and ethical behavior [5]. The nurses and social workers with ethics education use ethics resources more often [22]. In this way, ethics education is an important factor for nurses' moral perception and action. Therefore, we created a hypothesis that ethical education might change the perception of the ethical climate for nurses.

According to previous studies, there is no relationship between nurses' education degree and ethical climate score [11, 12, 15, 16, 23, 24]. However, continuing education includes post-license education; thus, it may be associated with nurses' perception of ethical climate. Grady posited access to continuing education as an indicator of the organization's ethical climate and support for ethics [22]. Actions for promoting a positive ethical climate include providing support for the continuing education on nursing ethics [12]. In the research regarding continuing education, Borhani et al. stated that the history of attendance to ethical workshops was related to a more positive evaluation of the ethical climate at the workplace [23], while other researchers insisted that there was no statistical difference between HECS and receiving training related to ethics [11] or ethics workshops [17]. The study of the employees in the executive branch of the U.S. federal government has shown that ethics training had a consistently positive influence on employee perception of ethical climate [25]. Moreover, in McDaniel's intervention study of critical care, the ethical environment score of nurses who received the ethics educational program was higher than that of nurses who had not [26].

https://doi.org/10.1371/journal.pone.0269034

Although these earlier studies investigated the relationship of several factors with ethical climate using the t-test, ANOVA, or single-liner model [11, 17, 24], adjusting for factors associated with ethical climate may yield different results. Therefore, in this report, we would like to more strictly focus on the relationship between nursing ethics education and ethical climate using multiple regression analysis, then elucidate the detailed educational contents that might influence nurses' perception of ethical climate. In Japan, the main types of continuing education for registered nurses (RNs) and midwives are in-service training in the workplace and workshops/ academic conferences. We have, hence, focused on these types of ethics education.

Methods

Design, Sample, and Settings

This study implemented a quantitative cross-sectional survey. The instrument used was a self-administered questionnaire. To the request for cooperation in the Kansai region, three teaching hospitals agreed to participate in this survey. These hospitals are teaching hospitals / academic medical centers that are large, multi-unit, acute-care general hospitals. They are in charge of assisting in the training of undergraduate nursing students and newly graduated nurses. Since the main purpose of this study was to explore the relationship between ethics education and ethical climate, we selected these teaching hospitals where the nursing staff is considered to have a high interest in education.

The questionnaires were delivered to the hospitals directly or through the mail, who then distributed them among the nurses who directly cared for patients. Upon completion, the questionnaires were collected by the researchers from the hospitals or mailed directly to the researcher by the nurses themselves. The survey was conducted from December 2019 to January 2020, with a population of 2168 nurses working in 12 units at three teaching hospitals. While 2168 questionnaires were distributed, 770 responses were received, returning a rate of 35.5%. Out of the total responses, 165 having missing data for J-HECS26 were excluded from the analyses. The sample size was considered sufficient to meet 351 nurses, which is calculated by the Cochran formula [27].

Measurement Tools

Among the various scales of evaluation of ethical climate, the HECS is used most to measure the ethical climate of the nursing environment [5]. The HECS was developed by Olson [4] in the U.S. to measure how nurses perceive the ethical climate of their workplace through their perspective of relationships with peers, patients, managers, hospital, and physicians. It consists of 26 items and five subscales: peers (4 items), patients (4 items), managers (6 items), hospital (6 items), and physicians (6 items). The survey was designed in a Likert-type format with all items being positive and answers rated from 1 (almost never true) to 5 (almost always true). The HECS has been examined for validity and reliability and is widely used in various countries [12, 13, 16, 28–31]. In Olson's study [4], the total value of Cronbach's alpha coefficient was found to be 0.91 and the values of the five subscales were as follows: peers– 0.73, patients– 0.68, managers– 0.81, hospital– 0.92, and physicians– 0.77. To focus on the ethical atmosphere in hospitals, we chose HECS to measure nurses' perceptions of ethical climate. In Japan, there are two types of translations of HECS, namely, J-HECS26 and J-HECS. The former, which we translated into Japanese, consists of all the 26 items in the original version, while the latter consists of 18 items, whose reliability and validity were confirmed in the study targeting the medium-sized hospitals with 180–500 beds [32]. The reliability and validity of J-HECS26 were verified in the study conducted in large teaching hospitals with 600–1100 beds [33]. The Cronbach's alpha coefficient of J-HECS26 in this study was 0.94 and their subscale are 0.67–0.95, which are equivalent to the original HECS in the previous literature [33].

https://doi.org/10.1371/journal.pone.0269034

Sociodemographic Characteristics

The questionnaire in the current study also encompassed the following demographic variables: age, gender, years of experience, marital status, education degree, the hospital nurses work at, license, position, the specialty of the unit they belong to, the experience of ethics education, and the opportunity of attending ethical training in and out of the hospitals (in-service training in the workplace, workshops / academic conferences on nursing ethics).

Regarding Ethics Education

In this research, we divided ethical education into two types; in-service training and the workshops/ academic conference. The in-service training on ethics in the workplace includes the ethics training sponsored by the hospital's Nursing Department and the study session on ethics held in the hospital ward. This type of training is intended to meet the actual needs of the workplace. The workshops/ academic conferences are held several times a year or annually by the nurses' associations or organizations. In the workshops/ academic conferences, researchers give research presentations or attend lectures on advanced studies for carrier development and continuous education. They are often held for academic purposes and the topics are usually general ethical problems.

Ethical Consideration

Ethical approval was obtained from the Research Ethics Committee of the Kyoto University (R0850). Each questionnaire had a cover letter explaining the purpose of the research. The identity of the participants was kept anonymous, and the return of the completed questionnaires was considered as evidence of informed consent for participation in the study. Each questionnaire was assigned a code number that helped mask the individual identities, and the responses from the nurses were kept confidential.

Statistical Analysis

To ensure the collection and analysis of data without compromising confidentiality, each questionnaire was treated anonymously, and each participant was indicated by a code number. The sample data were analyzed using the SPSS 24.0 J software for Windows (IBM Corporation). The mean and standard deviation for age and years of experience, and J-HECS26 and sub-scale items were calculated. As for the correlation between ethical climate and age, years of experience, a Pearson correlation was calculated. The means of J-HECS26 total score for the demographic variables of gender, marital status, education degree, hospital, unit (surgery, internal medicine, mixed surgery/internal medicine, psychiatry, obstetric, pediatrics, operating room, emergency department, stroke care unit (SCU), intensive care unit (ICU)/ coronary care unit (CCU), neonatal intensive care unit (NICU)), license (Midwife, Advanced licensed nurse), job position, and continuing education were evaluated using t-test, one-way analysis of variance (ANOVA), and Kruskal-Wallis test. Multivariate linear regression models were developed to analyze the relationship between the J-HECS26 and the independent variables: demographic variables and continuing education. Model 1 included the years of experience (continues), gender, job position, license (dummy variable), hospital (dummy variable), and 11 units (dummy variable). Model 2 included all the variables in Model 1, along with in-service training on ethics in the workplace. Model 3 included Model 2 variables plus workshops/ academic conferences on nursing ethics. The means of J-HECS26 score were compared between the data for nurses who have attended and those who did not attend in-service training in the workplace on ethics, and workshops/ academic conferences on nursing ethics, using analysis of covariance adjusted for year of experience (continues), gender, job position, license (dummy variable), hospital (dummy variable), and 11 units (dummy variable). The level of significance chosen for this study was .05.

https://doi.org/10.1371/journal.pone.0269034

Results

Demographics

The demographic data are shown in Table 8.3. Of the total 605 nurses, 561 (92.7%) who answered the questionnaire were females and 44 were males (7.3%) with a mean age of 31.4 ± 8.3 years. The mean number of years of their experience was 9.3 ± 8.0. Among the sample population, 167 (27.7%) had a nursing diploma, 369 (61.0%) had a bachelor's degree, and 20 (3.3%) had a master's / doctoral degree, while 48 (7.9%) were junior college graduates. Concerning ethics educational opportunities, 430 (71.1%) nurses had ethics education, while 175 (28.9%) nurses did not. Regarding the type pf education, 393 (65.0%) nurses attended in-service training at the workplace and 122 (20.2%) attended workshops/ academic conferences (Table 8.3).

Table 8.3 Demographic characteristics (n = 605).

Variables		n	(%)	Mean	SD
Age (year)				31.4	8.3
Years of experience (year)				9.3	8.0
Gender					
	Female	561	(92.7)		
	Male	44	(7.3)		
Education degree					
	Nursing diploma	167	(27.7)		
	Junior college	48	(7.9)		
	Bachelor's	369	(61.0)		
	Master's/doctoral	20	(3.3)		
	No answer	1	(0.2)		
Hospital					
	1	410	(67.7)		
	2	130	(21.6)		
	3	65	(10.7)		
License					
	Resisted Nurse	561	(92.7)		
	Midwife	29	(4.8)		
	Advanced Licensed Nurse	14	(2.4)		
	No answer	1	(0.1)		
Position					
	Staff	531	(87.8)		
	Manager	74	(12.2)		
Unit					
	Surgery	201	(33.3)		
	Internal medicine	174	(28.8)		
	Mixed surgery and internal medicine	69	(11.4)		
	ICU/CCU	35	(5.8)		
	NICU	28	(4.6)		

https://doi.org/10.1371/journal.pone.0269034

Variables		n	(%)	Mean	SD
	Obstetric	22	(3.6)		
	Pediatric	19	(3.1)		
	Operating room	16	(2.6)		
	Psychiatry	15	(2.5)		
	Emergency department	12	(2.0)		
	SCU	7	(1.2)		
	Others	7	(1.0)		
Experience of ethical education					
	Yes	430	(71.1)		
	No	175	(28.9)		
In-service training in workplace					
	Yes	393	(65.0)		
	No	212	(35.0)		
Workshops/ Academic conferences on nursing ethics					
	Yes	122	(20.2)		
	No	483	(79.8)		

Note; SD = standard deviation

https://doi.org/10.1371/journal.pone.0269034.t001

Nurses' Level of Ethical Climate

The mean score of J-HECS26 was 94.8 ± 14.8. The mean scores for the five sub-scales of the HECS, namely, managers, hospital, physicians, peers, and patients were 22.5 ± 5.9, 21.0 ± 3.6, 19.8 ± 4.2, 16.3 ± 2.5, and 15.2 ± 2.0, respectively (Table 8.4).

Table 8.4 J-HECS26 score (n = 605).

	Mean	SD	Min	Max
Total score of J-HECS26	94.8	14.8	47	130
Peers (4 items)	16.3	2.5	8	20
Patients (4 items)	15.2	2.0	8	20
Managers (6 items)	22.5	5.9	6	30
Hospital (6 items)	21.0	3.6	10	30
Physicians (6 items)	19.8	4.2	6	30

https://doi.org/10.1371/journal.pone.0269034.t002

Demographic Variables and J-HECS26

An independent sample correlation, t-test, and ANOVA were conducted to examine whether there is any relationship between J-HECS26 and the demographic variables. There was a weak negative correlation between J-HECS26 and age ($r = -0.130$, $p = 0.002$), years of experience ($r = -0.136$, $p = 0.001$). In the results of the t-test, one-way ANOVA, and Kruskal-Wallis test, a significant difference was found between J-HECS26 and demographic variables of the hospital ($p = 0.001$) and units ($p < 0.001$); notably, the score of nurses working in operating rooms was the lowest ($p < 0.001$) (Table 8.5). Female nurses scored higher

than males ($p = 0.042$) and the nurses who have had ethics education got a higher score than those who have not had any ($p = 0.035$). Regarding continuing education, nurses who have attended in-service training at the workplace scored higher in J-HECS26 than non-participants ($p < 0.001$). Further, the nurses who had attended workshops/ academic conferences on nursing ethics scored lower than those who had not ($p = 0.023$). The difference between J-HECS26 and demographic variables of marital status, education degree, license, and position were not statistically significant.

Table 8.5 Association between demographic variables and J-HECS26 total score (n = 605).

		n	Mean	SD	Min	Max	p value
Gender†							0.042*
	Female	561	95.1	14.8	47	130	
	Male	44	90.4	14.6	62	128	
Hospital‡							0.001*
	1	410	93.4	14.9	47	130	
	2	130	96.8	14.1	51	127	
	3	65	99.5	14.6	62	130	
License‡							0.174
	Registered Nurse	562	94.7	14.8	47	130	
	Midwife	29	98.6	14.2	71	127	
	Advanced Licensed Nurse	14	89.8	16.8	60	121	
Position†							0.558
	Manager	74	95.7	13.4	50	121	
	Staff	531	94.6	15.0	47	130	
Unit§ (n = 604)							< 0.001*
	Surgery**	201	92.9	14.8	47	122	
	Internal medicine**	174	97.3	12.0	63	128	
	Mixed surgery and internal medicine**	69	99.0	13.4	64	126	
	Psychiatry**	15	103.9	14.1	76	130	
	Pediatric**	19	95.8	19.2	53	127	
	Obstetric**	22	102.9	12.8	75	127	
	Emergency department**	12	98.6	15.7	72	124	
	Operating room	16	73.0	11.8	50	95	
	ICU/CCU	35	88.1	13.9	61	113	
	SCU	7	91.0	6.8	77	98	
	SCU NICU**	28	92.1	19.5	53	130	
	Others	7	87.7	14.5	71	116	
Experience of ethics education†							0.035*
	Yes	430	95.6	14.2	50	130	
	No	175	92.8	16.0	47	130	
In-service training in workplace on ethics†(n = 604)							< 0.001*
	Yes	393	96.3	14.0	50	130	
	No	211	91.8	15.8	47	130	

https://doi.org/10.1371/journal.pone.0269034

		n	Mean	SD	Min	Max	p value
Workshops/ Academic conferences on nursing ethics† (n = 604)							0.023*
	Yes	122	92.0	14.4	50	127	
	No	482	95.4	14.9	47	130	

Notes

† student' t-test

‡ one-way ANOVA: one-way analysis of variance

§ Kruskal-Wallis test; SD = standard deviation.

* $p < 0.05$

** Multiple comparison test $p < 0.05$ versus operating room

https://doi.org/10.1371/journal.pone.0269034.t003

Multiple regression analysis was used to assess the association between demographic variables and J-HECS26 (Table 8.6). The mean J-HECS26 score showed significant increase in manager nurses ($p = 0.033$) and nurses in the psychiatry unit ($p = 0.007$) obstetrics unit ($p = 0.014$), and emergency department ($p = 0.031$) and showed significant decrease in nurses with less experience ($p = 0.002$), male nurses ($p = 0.020$), and nurses in the operating room ($p = 0.004$).

Table 8.6 Results of multiple regression analysis with J-HECS26 (n = 604).

	Model 1†			Model 2‡			Model 3§		
	Coefficient	t	p value	Coefficient	t	p value	Coefficient	t	p value
Years of experience	-0.150	-3.1	0.002*	-0.151	-3.1	0.002*	-0.132	-2.7	0.007*
Gender	-0.091	-2.3	0.020*	-0.090	-2.3	0.022*	-0.089	-2.3	0.022*
Manager	0.105	2.1	0.033*	0.092	1.9	0.064	0.110	2.2	0.029*
Midwife	-0.073	-1.0	0.329	-0.076	-1.0	0.311	-0.072	-1.0	0.336
Advanced license nurse	-0.042	-1.0	0.298	-0.040	-1.0	0.327	-0.030	-0.7	0.465
Hospital 1	-0.152	-2.4	0.017*	-0.127	-2.0	0.049*	-0.116	-1.8	0.073
Hospital 2	-0.060	-1.0	0.339	-0.055	-0.9	0.383	-0.036	-0.6	0.569
Surgery	0.067	0.5	0.617	0.080	0.6	0.549	0.080	0.6	0.548
Internal medicine	0.210	1.6	0.110	0.220	1.7	0.093	0.223	1.7	0.088
Mixed surgery and internal medicine	0.143	1.5	0.138	0.151	1.6	0.117	0.155	1.6	0.106
Psychiatry	0.156	2.7	0.007*	0.159	2.8	0.006*	0.158	2.8	0.006*
Pediatric	0.066	1.1	0.284	0.060	1.0	0.331	0.063	1.0	0.303
Obstetrics	0.213	2.5	0.014*	0.214	2.5	0.013*	0.217	2.5	0.012*
Emergency department	0.106	2.2	0.031*	0.109	2.2	0.026*	0.107	2.2	0.029*
Operating room	-0.170	-2.9	0.004*	-0.162	-2.8	0.006*	-0.160	-2.7	0.007*
ICU/CCU	-0.026	-0.3	0.734	-0.023	-0.3	0.761	-0.019	-0.3	0.795

(*Continued*)

https://doi.org/10.1371/journal.pone.0269034

Table 8.6 Results of multiple regression analysis with J-HECS26 (n = 604). (*Continued*)

	Model 1†			Model 2‡			Model 3§		
SCU	0.000	0.0	0.996	0.005	0.1	0.913	0.006	0.1	0.906
NICU	0.022	0.3	0.762	0.029	0.4	0.681	0.031	0.4	0.661
In-service training in workplace on ethics				0.088	2.2	0.031*	0.089	2.2	0.028*
Workshops/ Academic conferences on nursing ethics							-0.092	-2.2	0.028*
R^2	0.156			0.164			0.171		

Notes

† Model1: Un-adjusted

‡ Model2: Adjusted for Model 1 covariates

§ Model3: Adjusted for Model 2 covariates.

* $p < 0.05$

https://doi.org/10.1371/journal.pone.0269034.t004

Continuing Education

Further multiple-regression analysis was conducted to assess the association between continuing education and J-HECS26 scores (Table 8.6). In Model 2, the mean J-HECS26 score showed a significant increase in nurses who had attended in-service training at the workplace ($p = 0.031$). In Model 3, the mean J-HECS26 score significantly decreased in nurses who had attended workshops/ academic conferences ($p = 0.028$).

Table 8.7 shows the multivariate-adjusted mean of the J-HECS26 score. The nurses who had attended in-service training at the workplace scored 2.63 points, which was significantly higher than those who had not attended any ($p = 0.038$, 95% CI: 0.15, 5.11). Nurses who had attended workshops/ academic conferences scored 3.30 points lower than those who did not ($p = 0.033$, 95% CI: - 6.33, - 0.28).

Table 8.7 Difference of multivariate-adjusted mean values of J-HECS26 (n = 604).

	(Yes)			(No)			(Yes)- (No)			
	Mean	95% CI		Mean	95% CI		Mean	95% CI		p value
In-service training in workplace on ethics	95.7	94.3	97.0	93.0	91.0	95.0	2.63	0.15	5.11	0.038*
Workshops/ Academic conferences on nursing ethics	92.1	89.4	94.8	95.4	94.1	96.7	-3.30	-6.33	-0.28	0.033*

Notes: Adjusted for year of experience (continues), gender, hospital (dummy variables), manager, unit (surgery, internal medicine, mixed surgery and internal medicine, psychiatry, pediatric, obstetric, emergency department, operation, ICU/CCU, SCU, NICU: dummy variables), license (midwife, advanced license nurse: dummy variables); CI = confidence interval.

* $p < 0.05$

https://doi.org/10.1371/journal.pone.0269034.t005

Discussion

The purpose of this study was to ascertain the factors that might affect the ethical climate by using J-HECS26, which we first translated into Japanese before distributing it among the participants. We focused on the demographic data and education since the detailed influence of education on HECS has not been fully reported so far.

Nurses' Level of Ethical Climate

The present study indicated that the mean score of J-HECS26 and sub-scales were 94.8. The mean score of J-HECS26 in this study was more than that in Khalesi [34] (71.5) and Shafipour, Yaghobian, Shafipour, and Heidari [35] (91.8), whereas it was less than that in Bansal [16] (116.7) and Tehranineshat [36] (100.1). All these studies were conducted in teaching hospitals, as was done in our study. It was found that nurses in this study perceived their ethical climate as relatively desirable.

Demographic Characteristics

There was a significant relationship between nurses' demographic characteristics and their perception of ethical climate. Younger nurses and those with less experience have a more positive perception of ethical climate. Being a female was related to a more positive evaluation of the ethical climate of the hospitals where they were employed. Additionally, nurses' perception of ethical climate was significantly related to the hospital and unit where they worked. There are several previous studies regarding factors influencing the HECS, although some are rather controversial. Some studies reported that age, gender, years of experience, and degree of education (i.e., nursing diploma /associate /bachelor's /master's degree) had no significant impact on the HECS scores [16, 18, 37], whereas other studies showed that age [12, 15], gender [11, 24], job position [12], years of experience [11, 12, 13, 15] and education degree (bachelor's /master's /doctoral degree) did have a statistically significant effect on the HECS [17].

Concerning years of experience, inexperienced nurses were reported to perceive ethical climate optimistically [13, 15], which is consistent with our results in this current study. Concerning gender, similar to our findings, Ozden et al. reported a significant difference between the gender of nurses and ethical climate perceptions, claiming that their result reflected the general situation of the majority of participants in the survey being female, as Turkish nurses were predominantly female [11]. In our study, 92% of the participants were female, which is similar to the overall proportion of nurses in Japan [38]. There are several reports that male nurses feel more difficulty with working in a female-centered workplace as compared to female nurses in terms of interacting with female patients and female nurses and anxiety regarding their future career development [39]. Therefore, male nurses might recognize their perception of ethical climate more negatively than female nurses.

Further, the t-test results showed that there was no significant statistical difference between the perception of the ethical climate of manager nurses and staff nurses. However, adjusting for variables such as years of experience and gender, managers have significantly higher perceptions of ethical climate. In previous literature, Hwang and Park reported that manager nurses perceive ethical climate more positively than staff nurses [12]. Our study is in accordance with their study, demonstrating a similar result even after adjusting for years of experience. These studies, including ours, indicate that the managers might generally commit to the ethical procedures and policies of the organization and collaborate with doctors and colleagues on ethical issues.

Regarding the hospital and specialty of the unit, previous studies suggested that the staff in different hospitals had different perceptions of the ethical climate [11, 12, 15]. Ozden et al.'s study [11], which was conducted in a university-attached hospital and an education and research hospital, similar to our study, reported a significant difference in the ethical climate among hospitals. Our study supports their report.

https://doi.org/10.1371/journal.pone.0269034

On the other hand, a review of the related studies presents conflicting results regarding the relationships between the unit and nurses' perceptions of ethical climate. Hwang and Park [12] argued that the HECS scores were not related to the specialties of the units, while Ozden et al. [11] and Lemmenes et al. [40] confirmed that they were indeed related to this variable. This inconsistency might be relevant to the differences between the researched hospitals. Our results showed that the ethical climate differs among the units; in particular, nurses in the operating room showed significantly lower mean scores of J-HECS26. Rosenstein and O'Daniel noted that the operating room is physically small, there is a strong interdependence on effective team functioning, and working in this unit is highly stressful [41]. According to Sonoda, Onozuka, and Hagihara, operating room nurses have also been reported to be mentally stressed not only by doctors but also by colleagues [42]. Circulating nurses in the operation room may feel frustrated by the scrub nurses [42], presumably leading to a negative perception of the workplace. The low mean score of nurses in the operating room in our study might reflect a negative relationship with these colleagues in the operation room or at the workplace.

Continuing Education

Other than the aforementioned demographic characteristics, we found that ethics education may have influenced the nurses' perception of the ethical climate. This study showed that nurses who had attended in-service training at the workplace perceived their ethical climate positively; however, the nurses who had participated in the workshops/ academic conferences found their hospital ethical climate to be negative, which differed from our expectations. This result was verified through multiple regression analysis and analysis of covariance.

Previous studies showed conflicting results regarding the effect of ethics education on the HECS scores. While some studies showed no relationship between ethical climate and attendance of ethical workshop/ training [11, 17], Borhani et al. reported a positive effect of education on ethical climate, indicating that the nurses who participated in an ethics workshop perceived their ethical climate as more positive [23]. These differences might depend on the style of education or programs. Previous literature pointed out that interactive workshops, rather than uni-directional lectures, enlighten nurses more [43]. An ethical climate represents "shared perceptions" of organizational practices related to ethical decision-making [3]. In-service training in the workplace gives nurses a consistent awareness of the ethics practiced in the particular hospital where they are employed. As a result, nurses who attend in-service training may be more positive about the ethical climate of their workplace than those who did not.

On the other hand, nurses who participated in off-the-job workshops/ academic conferences perceive their climate more negatively than those who did not participate. At workshops/ academic conferences, nurses acquire the most advanced ethical knowledge, values, and perception. However, it is unclear whether this knowledge will be properly utilized in the nursing practice of their workplace. In fact, one study shows that off-the-job training may not directly apply to the workplace of nurses [44]. Therefore, what they have learned there may not have been relevant to their practical situation; thus, they may not have been able to apply what they have learned or share it with their colleagues. These frustrations might explain the negative relationship between ethical climate and the out-of-hospital ethical training found in this study. Therefore, what they have learned at off-the-job workshops/ conferences should be reviewed and incorporated into their workplace with caution to ensure that their acquired knowledge finds more use in practice.

Additionally, having education outside the hospital is not enough to enhance the workplace's ethical climate. Ethical climate is nurses' shared perception. Even if a nurse adapts the latest knowledge at an academic conference or out-of-hospital educational opportunities, it may be difficult to put it into practice if it is not properly shared with colleagues. On the contrary, the same knowledge is shared by the nurses of the same hospital at in-hospital training, which can transform the acquired knowledge into practical use. Ulrich also states that receiving ethical education alone may leave nurses dissatisfied unless it is properly practiced in the workplace [45]. This supports our result that in-hospital training is more related to the positive ethical climate.

https://doi.org/10.1371/journal.pone.0269034

Moreover, the motivation for nurses to receive continuing professional education is to improve knowledge, patient care delivery, and professional relationships [46]. There is a possibility that the nurses who had attended the out-of-hospital education had negative feelings about their current workplace, involving concerns related to ethical issues and seeking solutions. Accordingly, the nurses who had attended the out-of-hospital ethical education might have a negative perception of the ethical climate of their workplace in this study. Resolving the frustrations and concerns of these nurses may lead to reducing their moral distress and preventing their intention to leave the organization. Simultaneously, the HECS may be useful in identifying the nurses who need support from the workplace to make the best of their eagerness to improve the ethical atmosphere.

In summary, our study has new implications for the importance of the provision of ethics education and clarifies that offering in-service ethical training that meets the actual needs of their workplace may provide nurses a shared perception of ethical climate. Our result demonstrates that not only individual efforts but also organizational efforts are essential for improving the ethical climate of a workplace. Further research should elucidate the kind of education and organizational support that is effective in enhancing the ethical climate.

Strength and Limitations

The strength of this study is that it is the first to use J-HECS26 (translated into Japanese), with all 26 items of the original HECS. Further, our study clearly demonstrated a relationship between demographic characteristics and ethics education by using multiple regression analysis. However, our study did not clarify the details of ethics education (i.e., timing, frequency, and content). Moreover, since educational support differs depending on the hospital, a selection bias may have occurred. Therefore, future studies must analyze hospitals of other sizes and that offer different services in order to clarify the details of education. Ethical climate has been reported to be related to various other factors such as ethical stress or job satisfaction. Thus, another limitation is that these factors were not measured in our questionnaire.

Conclusion

This study examined nurses' perceptions of the ethical climate in Japanese teaching hospitals. Nurses' perception of ethical climate varied based on individual characteristics such as years of experience, gender, job position, hospital, and specialty of the unit. We also demonstrated that in-service training, adjusted for the workplace, is effective in raising employee perception of the ethical climate. Lastly, off-the-job workshops/ academic conferences are not directly related to enhancing the ethical climate and, hence, should be implemented carefully.

Supporting Information

S1 Dataset.
(XLSX)

Acknowledgments

The authors are grateful to all the participants for their contribution to the study.

Author Contributions

Conceptualization: Ayaka Okumoto, Chiharu Miyata.
Data curation: Ayaka Okumoto, Satoko Yoneyama.
Formal analysis: Ayaka Okumoto, Satoko Yoneyama.
Funding acquisition: Ayaka Okumoto, Ayae Kinoshita.

https://doi.org/10.1371/journal.pone.0269034

Investigation: Ayaka Okumoto, Ayae Kinoshita.
Methodology: Ayaka Okumoto, Satoko Yoneyama, Ayae Kinoshita.
Project administration: Ayaka Okumoto, Ayae Kinoshita.
Resources: Ayaka Okumoto.
Supervision: Satoko Yoneyama, Chiharu Miyata, Ayae Kinoshita.
Writing – original draft: Ayaka Okumoto.
Writing – review & editing: Ayaka Okumoto, Satoko Yoneyama, Chiharu Miyata, Ayae Kinoshita.

References

1. Olson LL. Envisioning an Ethical Climate in Nursing Education Programs. Online Journal of Issues in Nursing. 2021;26(1):1–10. https://doi.org/10.3912/OJIN.Vol26No01Man07
2. Victor B, Cullen JB. The organizational bases of ethical work climates. Administrative Science Quarterly. 1988:101–25. https://doi.org/10.2307/2392857
3. Olson LL. Ethical climate in health care organizations. International nursing review. 1995;42(3) :85–90. PMID:7649724
4. Olson LL. Hospital nurses' perceptions of the ethical climate of their work setting. Image: the Journal of Nursing Scholarship. 1998;30(4):345–9. https://doi.org/10.1111/j.1547-5069.1998.tb0 1331.x PMID:9866295
5. Koskenvuori J, Numminen O, Suhonen R. Ethical climate in nursing environment: a scoping review. Nursing ethics. 2019;26(2):327–45. https://doi.org/10.1177/0969733017712081 PMID:28659068
6. Schluter J, Winch S, Holzhauser K, Henderson A. Nurses' moral sensitivity and hospital ethical climate: A literature review. Nursing ethics. 2008;15(3):304–321. https://doi.org/10.1177/096973 3007088357 PMID:18388166
7. Hamric AB, Blackhall LJ. Nurse-physician perspectives on the care of dying patients in intensive care units: collaboration, moral distress, and ethical climate. Critical Care Medicine. 2007;35(2):422–9. https://doi.org/10.1097/01.CCM.0000254722.50608.2D PMID:17205001
8. Silén M, Svantesson M, Kjellström S, Sidenvall B, Christensson L. Moral distress and ethical climate in a Swedish nursing context: perceptions and instrument usability. Journal of clinical nursing. 2011;20(23-24):3483–93. https://doi.org/10.1111/j.1365-2702.2011.03753.x PMID:21884556
9. Parker FM, Lazenby RB, Brown JL. The relationship of moral distress, ethical environment and nurse job satisfaction. Online Journal of Health Ethics. 2013;10(1):2. http://dx.doi.org/10.18785/ojhe.1001.02
10. Joseph J, Deshpande SP. The impact of ethical climate on job satisfaction of nurses. Health care management review. 1997;22(1):76–81. PMID:9058089
11. Özden D, Arslan GG, Ertuğrul B, Karakaya S. The effect of nurses' ethical leadership and ethical climate perceptions on job satisfaction. Nursing ethics. 2019;26(4):1211–25. https://doi.org/10.11 77/0969733017736924 PMID:29117776
12. Hwang J-I, Park H-A. Nurses' perception of ethical climate, medical error experience and intent-to-leave. Nursing ethics. 2014;21(1):28–42. https://doi.org/10.1177/0969733013486797 PMID:23793070

13. Lützen K, Blom T, Ewalds-Kvist B, Winch S. Moral stress, moral climate and moral sensitivity among psychiatric professionals. Nursing Ethics. 2010;17(2):213–24. https://doi.org/10.1177/0969733009351951 PMID:20185445

14. Cerit B, Özveren H. Effect of hospital ethical climate on the nurses' moral sensitivity. The European Research Journal. 2019;5(2):282–90. https://doi.org/10.18621/eurj.423324

15. Bahcecik N, Oztürk H. The hospital ethical climate survey in Turkey. JONA'S healthcare law, ethics and regulation. 2003;5(4):94–9. https://doi.org/10.1097/00128488-200312000-00008 PMID:14660940

16. Bansal AK, Parmar PB, Bansal P, Patel R, Barai PH, Thomas E. Ethical climate and its effect in teaching hospital: A vision from 3rd eye. Journal of Indian Academy of Forensic Medicine. 2019;41(1):45–9. https://doi.org/10.5958/0974-0848.2019.00011.3

17. Gilvari T, Abbaszadeh A, Borhani F, Mohamadi P, Saberi N. Relationship of the Hospital Ethical Climate with Nurses' Attitude to Interprofessional Collaboration. Journal of Clinical & Diagnostic Research. 2019;13(11). https://doi.org/10.7860/JCDR/2019/42752.13324

18. Pauly B, Varcoe C, Storch J, Newton L. Registered nurses' perceptions of moral distress and ethical climate. Nursing ethics. 2009;16(5):561–73. https://doi.org/10.1177/0969733009106649 PMID:19671643

19. Bell SE. Ethical climate in managed care organizations. Nursing administration quarterly. 2003;27(2):133–9. https://doi.org/10.1097/00006216-200304000-00007 PMID:12765105

20. Okumoto A, Miyata C, Yoneyama S, Kinoshita A. Nurses' perception of the bed alarm system in acute-care hospitals. SAGE Open Nursing. 2020;6:1–10. https://doi.org/10.1177/2377960820916252 PMID:33415274

21. Corley MC. Nurse moral distress: a proposed theory and research agenda. Nursing ethics. 2020;9(6):636–650.

22. Grady C, Danis M, Soeken KL, O'Donnell P, Taylor C, Farrar A, et al. Does ethics education influence the moral action of practicing nurses and social workers? The American Journal of Bioethics. 2008;8(4):4–11. https://doi.org/10.1080/15265160802166017 PMID:18576241

23. Borhani F, Abbaszadeh A, Bahrampour A, Ameri GF, Aryaeenezhad A. Investigating the relationship between the ethical atmosphere of the hospital and the ethical behavior of Iranian nurses. Journal of Education and Health Promotion. 2021;10(1):193. https://doi.org/10.4103/jehp.jehp_891_20 PMID:34250127

24. Abou Hashish EA. Relationship between ethical work climate and nurses' perception of organizational support, commitment, job satisfaction and turnover intent. Nursing ethics. 2017;24(2):151–66. https://doi.org/10.1177/0969733015594667 PMID:26260440

25. Raile ED. Building ethical capital: Perceptions of ethical climate in the public sector. Public Administration Review. 2013;73(2):253–62. https://doi.org/10.1111/j.1540-6210.2012.02649.x

26. McDaniel C. Ethical environment: reports of practicing nurses. The Nursing Clinics of North America. 1998;33(2):363–72. PMID:9624210

27. Kotrlik J, Higgins C. Organizational research: Determining appropriate sample size in survey research appropriate sample size in survey research. Information technology, learning, and performance journal. 2001;19(1):43.

28. Asgari S, Shafipour V, Taraghi Z, Yazdani-Charati J. Relationship between moral distress and ethical climate with job satisfaction in nurses. Nursing Ethics. 2019;26(2):346–56. https://doi.org/10.1177/0969733017712083 PMID:28718349

29. Charalambous A, Cloconi C, Papastavrou E, Theodoula A. Psychometric properties of the hospital ethical climate survey: A cross-sectional study in Greek and cypriot cancer care settings. Journal of nursing measurement. 2018;26(2):237–48. https://doi.org/10.1891/1061-3749.26.2.237 PMID:30567942

30. Claeys M, Faelens A, Sabbe B, Schrijvers D, de Casterle BD, Luyten P. Psychometric properties of the Hospital Ethical Climate Survey: a cross-sectional study in a large sample of Belgian psychiatric nurses. European Journal for Person Centered Healthcare. 2013;1(1):202–8. http://dx.doi.org/10.5750/ejpch.v1i1.652

31. Suhonen R, Stolt M, Gustafsson ML, Katajisto J, Charalambous A. The associations among the ethical climate, the professional practice environment and individualized care in care settings for older people. Journal of advanced nursing. 2014;70(6):1356–68. https://doi.org/10.1111/jan.12297 PMID:24251452

32. Inagaki S, Osawa A, Yoshikawa A, Ishihara I. Development and validation of the Japanese version of the Hospital Ethical Climate Survey. The Japan Nursing Ethics Association. 2020; 12(1): 73–79.

33. Okumoto A, Miyata C, Yoneyama S, Kinoshita A. The Reliability and Validity of Japanese Version of Hospital Ethical Climate Survey. Journal of Japan Academy of Nursing Science. 2021;41:647–55. https://doi.org/10.5630/jans.41.647

34. Khalesi N, Arabloo J, Khosravizadeh O, Taghizadeh S, Heyrani A, Ebrahimian A. Psychometric properties of the Persian version of the "Hospital Ethical Climate Survey". Journal of medical ethics and history of medicine. 2014;7.

35. Shafipour V, Yaghobian M, Shafipour L, Heidari MR. Nurses' perception of the ethical climate in the Iranian hospital environment. Journal of Nursing and Midwifery Sciences. 2016;3(4):37–43. https://doi.org/10.18869/acadpub.jnms.3.4.37

36. Tehranineshat B, Torabizadeh C, Bijani M. A study of the relationship between professional values and ethical climate and nurses' professional quality of life in Iran. International journal of nursing sciences. 2020;7(3):313–9. https://doi.org/10.1016/j.ijnss.2020.06.001 PMID:32817854

37. Ghorbani AA, Hesamzadeh A, Khademloo M, Khalili S, Hesamzadeh S, Berger V. Public and private hospital nurses' perceptions of the ethical climate in their work settings, Sari City, 2011. Nursing and midwifery studies. 2014;3(1). https://doi.org/10.17795/nmsjournal12867 PMID:25414890

38. Japanese Nursing Association. Statistical Data on Nursing Service in Japan. [Internet]. [cited 2021 Nov 1]. Available from: https://www.nurse.or.jp/jna/english/statistics/

39. Takahashi R., Tanaka M., & Nin K. Sources and Coping mechanisms of Work-Induced Stress for Male Nurses. Annual reports of School of Health Sciences Faculty of Medicine, Kyoto University: health science. 2014; 9: 41–51. https://doi.org/10.14989/185393

40. Lemmenes D, Valentine P, Gwizdalski P, Vincent C, Liao C. Nurses' perception of ethical climate at a large academic medical center. Nursing ethics. 2018;25(6):724–33. https://doi.org/10.1177/0969733016664980 PMID:27605558

41. Rosenstein AH, O'Daniel M. Impact and implications of disruptive behavior in the perioperative arena. Journal of the American College of Surgeons. 2006;203(1):96–105. https://doi.org/10.1016/j.jamcollsurg.2006.03.027 PMID:16798492

42. Sonoda Y, Onozuka D, Hagihara A. Factors related to teamwork prformance and stress of operating room nurses. Journal of nursing management. 2018;26(1):66–73. https://doi.org/10.1111/jonm.12522 PMID:28744975

43. Griscti O, Jacono J. Effectiveness of continuing education programmes in nursing: literature review. Journal of Advanced nursing. 2006;55(4):449–56. https://doi.org/10.1111/j.1365-2648.2006.03940.x PMID:16866840

44. Vasanthi S, Basariya S. Pros and Cons of On the Job Training Versus Off the Job Training. International Journal of Scientific and Technology Research. 2019;8(10):671–4.

45. Ulrich C, O'Donnell P, Taylor C, Farrar A, Danis M, & Grady C. Ethical climate, ethics stress, and the job satisfaction of nurses and social workers in the United States. Social science & medicine. 2007;65(8):1708–19. https://doi.org/10.1016/j.socscimed.2007.05.05043

46. Smith J, Topping A. Unpacking the value added impact of continuing professional education: a multi-method case study approach. Nurse Education Today. 2001;21(5):341–9. https://doi.org/10.1054/nedt.2001.0564 PMID:11403580

8.9 Case Study 3

RESEARCH ARTICLE

Personal ethical settings for driverless cars and the utility paradox: An ethical analysis of public attitudes in UK and Japan

Kazuya Takaguchi[1,2], Andreas Kappes [3], James M. Yearsley[3], Tsutomu Sawai [4,5], Dominic J. C. Wilkinson [1,6,7]*, Julian Savulescu[1,8]

1 Oxford Uehiro Centre for Practical Ethics, Faculty of Philosophy, University of Oxford, Oxford, United Kingdom, 2 The Department of Ethics, Kyoto University, Kyoto, Japan, 3 City, University of London, London, United Kingdom, 4 Graduate School of Humanities and Social Sciences, Hiroshima University, Hiroshima, Japan, 5 Institute for the Advanced Study of Human Biology (ASHBi), Kyoto University, Kyoto, Japan, 6 John Radcliffe Hospital, Oxford, United Kingdom, 7 Murdoch Children's Research Institute, Melbourne, Australia, 8 Melbourne Law School, Melbourne, Australia

* *dominic.wilkinson@philosophy.ox.ac.uk*

Abstract

Driverless cars are predicted to dramatically reduce collisions and casualties on the roads. However, there has been controversy about how they should be programmed to respond in the event of an unavoidable collision. Should they aim to save the most lives, prioritise the lives of pedestrians, or occupants of the vehicle? Some have argued that driverless cars should all be programmed to minimise total casualties. While this would appear to have wide international public support, previous work has also suggested regional variation and public reluctance to purchase driverless cars with such a mandated ethical setting. The possibility that algorithms designed to minimise collision fatalities would lead to reduced consumer uptake of driverless cars and thereby to higher overall road deaths, represents a potential "utility paradox". To investigate this paradox further, we examined the views of the general public about driverless cars in two online surveys in the UK and Japan, examining the influence of choice of a "personal ethical setting" as well as of framing on hypothetical purchase decisions. The personal ethical setting would allow respondents to choose between a programme which would save the most lives, save occupants or save pedestrians. We found striking differences between UK and Japanese respondents. While a majority of UK respondents wished to buy driverless cars that prioritise the most lives or their family members' lives, Japanese survey participants preferred to save pedestrians. We observed reduced willingness to purchase driverless cars with a mandated ethical setting (compared to offering choice) in both countries. It appears that the public values relevant to programming of driverless cars differ between UK and Japan. The highest uptake of driverless cars in both countries can be achieved by providing a personal ethical setting. Since uptake of driverless cars (rather than specific algorithm used) is potentially the biggest factor in reducing in traffic related accidents, providing some choice of ethical settings may be optimal for driverless cars according to a range of plausible ethical theories.

https://doi.org/10.1371/journal.pone.0275812 Used with CC 4.0 license

1. Introduction

1.1 Driverless cars

Advances in motor vehicle design have reduced the devastating harm associated with traffic collisions. For example it was estimated that forward-collision warning and autonomous braking system prevented about 14% of crash fatalities in 2016 in the US [1]. However, as many as 1.35 million deaths and 50 million injuries still occur every year worldwide from road traffic accidents [2]. This is the 8th highest cause of death, and the leading cause of death for young people aged 5–29. Further technological advances, particularly the advent of automated driving technology (driverless cars) could dramatically reduce this. In Germany, the US and UK, between 90–95% of car accidents are estimated to be caused by human error or misconduct (for example, speeding, inattention, failing to give way) [3–5]. Due to the high rates of accidents caused by human error, driverless cars are believed to have a positive impact on road safety. For example, one study estimated up to 73% reduction in pedestrian crashes in Finland [6], while a US study estimated up to 90% reduction [7]. Another survey suggests that if all human-driven cars were replaced by fully automated driverless cars this could in theory prevent 30,000 lives per year in the US [8]. Considering the great reduction in number of injuries and fatalities, this would be a massive benefit for road users [9] and the wider community by saving social expenditure [10]. These advantages have motivated governments worldwide to facilitate the adoption of driverless cars, and car manufacturers and leading tech companies to compete in their development [11].

However, the development of fully automated cars raises a number of ethical questions. One such question is how such cars should be programmed to respond in the event of a collision. Although automation will potentially eradicate human error, accidents will continue to occur (for example, due to environmental factors, technological failure, or following unexpected behaviours by other road users) albeit with reduced frequency. Faced with an imminent, unavoidable collision, human drivers have limited time or ability to respond. In contrast, fully automated vehicles can be pre-programmed to respond in one or more specific ways taking into account information from the environment and context of the accident. For example, a driverless car detecting an imminent crash could seek to protect occupants of the vehicle, it could seek to avoid harming pedestrians (or other innocent "bystanders" such as cyclists), or it could aim to minimise overall casualties (save the most lives).

1.2 Ethical Programming of Driverless Cars

These different programming alternatives have been debated from the point of view of ethical theory. For example, utilitarian or consequentialist approaches typically support decisions that would minimise overall numbers of deaths or injuries (i.e., save the most lives). Other approaches (often drawing on variations of the philosophical thought experiment "the trolley problem" [12]) have questioned the idea that it would be ethical to deliberately direct a vehicle in a way that would kill pedestrians or other innocent parties. The German Ethics Commission produced a report indicating (rule 9) that driverless cars should *not* sacrifice pedestrians to save occupants [13]. On the other hand, prospective users of driverless cars may wish to protect themselves and other passengers (particularly if the passengers are friends or family members). There is a question about whether such partiality to occupants of the vehicle could be justified.

In an attempt to help guide the development of driverless cars, some empirical work has examined the views of the general public on how they think such cars should respond to collisions (their "Moral Algorithm Preference"). For example, in the Moral Machine Experiment (MME), online responses were obtained from almost 40 million decisions from participants in more than 200 countries [14]. Respondents judged a series of hypothetical collisions. The strongest preferences were for sparing humans over non-human animals, saving more lives, and sparing the young over the old [14]. Other studies appear to confirm

https://doi.org/10.1371/journal.pone.0275812

widespread support for saving the most lives in a collision (though often choosing to save pedestrians if there are similar numbers of occupants/pedestrians at risk) (Table 8.8). A follow-up study to the MME found that about 40% of respondents chose to treat groups of potential accident casualties equally when given that option (rather than saving the greatest number) [15]. Other studies have explored the way that question framing influences responses [16–18]. When instructed to adopt the perspective of a pedestrian, respondents were more likely to support driverless car responses that would endanger occupants rather than themselves [17, 18]. However, overall these studies suggest that the public may find it acceptable for driverless cars to be programmed with algorithms designed to minimise overall casualties, leading some to suggest that cars should be programmed with an algorithm saving the most lives as a mandatory ethical setting taking this into account [9].

Table 8.8 Studies investigating views of the public about ethical response to collision scenarios.

Author	Type of study	Style of study (methodology)	Participants	Location	Dominant preference
Awad et al., 2018 [14]	Moral Algorithm Preference	Online survey	>500,000 (40 million decisions)	233 countries	'Save the Most', (but some global variation)
					Some preference for pedestrians over occupants
Awad et al., 2020 [21]	Moral Algorithm Preference	Online survey	585,531 (alternative part of MME)	233 countries	'Save the Most', no preference for protecting passengers
Bergmann et al., 2018 [22]	Moral Algorithm Preference	Virtual reality simulation	189	Germany	'Save the Most'
Bigman and Gray, 2020 [15]	Moral Algorithm Preference	Online survey	2352+843 +993	US, UK	'Save the Most' (but 40% chose to treat equally)
Bonnefon et al., 2016 [19]	Moral Algorithm Preference	Online survey	1928 (total)	US	'Save the Most'
	Purchase Preferences				'Save the Occupants'
Faulhaber et al., 2019 [23]	Moral Algorithm Preference	Virtual reality simulation	189	Germany	'Save the Most'
Frank et al., 2019 [16]	Moral Algorithm Preference	Online survey	12,000	US, Denmark	Save pedestrian if low numbers in car (1, or 2), Otherwise 'Save the Most'
					(some influence of framing/perspective)
Kallionen et al., 2019 [17]	Moral Algorithm Preference (virtual reality/ animation)	Virtual reality simulation + online survey	184 + 368	Germany	'Save the Most'
					(some influence of framing)
Li et al., 2019 [24]	Moral Algorithm Preference	Virtual reality simulation	60	China	'Save the Most'

https://doi.org/10.1371/journal.pone.0275812

Author	Type of study	Style of study (methodology)	Participants	Location	Dominant preference
Liu, P and Liu, J, 2021 [20]	Moral Algorithm Preference	Online survey	580	China	No preference for 'Save the Most' vs 'Save the Occupants'
	Purchase Preferences				Willingness to pay is higher with 'Save the Occupants'
Mayer et al., 2021 [18]	Moral Algorithm Preference	Online survey	1380	Germany	Save pedestrian if equal numbers, otherwise 'Save the Most'
Pugnetti and Schläpfer 2018 [25]	Moral Algorithm Preference	Online survey	107	Swiss	'Save the Most'. Equal preference for pedestrians/occupants.
Wintersberger et al., 2017 [26]	Moral Algorithm Preference (driving simulator)	Driving simulator	40	Germany	'Save the Most'
					Survival rate influences their preference. From occupants' perspective, higher personal survival rates motivate to choose altruistic decisions

* The table indicates preferences for how respondents think driverless cars should respond (or the choice they personally would make) in the event of a collision (Moral Algorithm Preference), and which car they would actually purchase (Purchase Preference).
**We searched The National Centre for Biotechnology Information PubMed in March 2022 for papers using the keywords ("Autonomous" OR "self-driving" or "driverless") AND ("public opinion" OR "preference") AND ethics. Additional papers were identified from reference lists, related papers and the authors' libraries. Only studies reporting trade-off scenarios/preferences between saving larger/smaller numbers of people, and between saving occupants vs pedestrians are included.

https://doi.org/10.1371/journal.pone.0275812.t001

One challenge is that there may be significant variations between communities in the values that they would apply to driverless car collisions. In the MME study, respondents from a "Western" cluster of countries (including North America and many European countries) had a much stronger preference for saving the most lives than respondents from "Eastern" countries (including Japan, Taiwan, China, India, and many Middle Eastern countries) [14]. The Eastern cluster had a stronger preference for saving pedestrians. Thus, the above suggestion may not be appropriate to a country where public moral preferences are different.

1.3 Utility Paradox

A further challenge is how the public's views about the moral acceptability of driverless car programming would translate into their actual behaviour. Bonnefon et al. investigated not only public moral intuition about different driverless car algorithms, but also US consumers' willingness to buy those driverless cars ("Purchase Preferences"). Three quarters of their participants indicated that driverless cars should be programmed to minimise the number of victims (preferences were particularly strong where this would save

multiple lives such as 10 people) [19]. This tendency remained (though was weaker) when participants were asked to imagine a family member in the vehicle. However, when participants were asked about purchasing driverless cars, a strikingly different response was obtained–with low support for actually buying a utilitarian-programmed vehicle. This hesitation was even greater when participants were asked to imagine that their family members might use their driverless cars. Bonnefon and colleagues concluded that although people share moral intuitions supporting utilitarian algorithms, they do not want to buy such driverless cars because of concerns for their and their family's safety [19]. Likewise, a study of 580 Chinese participants indicated a preference for purchasing driverless cars that would save occupants over pedestrians, including a willingness to pay more for such a vehicle [20] (Table 8.8).

This suggests a potential *utility paradox* for driverless cars: driverless car algorithms that are designed to minimise collision fatalities may lead to reduced consumer uptake, higher use of non-autonomous cars and higher overall road deaths. Purchase choices may be particularly important to consider because of the very large potential difference in risk of fatal accidents with driverless cars (compared to non-autonomous vehicles).

Because of the significant potential benefit to the community by introducing driverless cars, it would be important to assess which approach would lead to the greatest uptake. It may be preferable to either adopt a different mandatory ethical setting (i.e. other than 'save the most lives') or to allow prospective consumers to choose which they would prefer (a so-called "personal ethical setting" [9]). Empirical research is needed that would compare the expected uptake rates of different forms of mandatory ethical setting, with that of a personal ethical setting. As purchase choices may be sensitive to the price of the case we also need to assess consumers' willingness to pay for their preferred algorithm. No previous studies to our knowledge have evaluated how providing a choice of driverless car algorithm would influence expected uptake of driverless cars, nor how this might vary between cultures.

1.4 This Study

We aimed to study how the potential consumers in two countries believe driverless cars should be programmed to respond in collisions (we call this "Moral Algorithm Preference"), as well as their hypothetical willingness to purchase these cars (we call this "Purchase Preference"). We chose to compare the UK with Japan, since previous research suggests different ethical values impacting on driverless car preferences. We compared purchasing preferences for different programming algorithms (whether mandated or provided as an option) and examined how these choices are influenced by framing effects and purchase price. Based on these data of purchasing behaviours, we attempt to seek what algorithm or which choice of algorithm would lead to the highest driverless car uptake if made available.

2. Methods

2.1 Participants

The UK Survey was conducted online from April 7 to May 15, 2021. UK residents aged over 18 were recruited via Prolific, using convenience sampling [27]. We performed the same survey with Japanese participants from December 6, 2021, to January 5, 2022, using Crowdworks (crowdworks.jp), again using convenience sampling. This project received ethical approval from the University of Oxford (the Society and Humanities Interdivisional Research Ethics Committee, IDREC9).

For the UK survey, using G*Power [28], we calculated that a sample size of 200 participants would have sufficient power (>80%) to detect small to medium differences in price sensitivity between the driverless car models. Crowdworks has an online Japanese worker pool of 4.1 million and has been previously validated for

https://doi.org/10.1371/journal.pone.0275812

psychology experiments [29]. We sought a larger sample (300) in Japan to take into account the possibility of higher drop-out with the broad online pool of freelance workers. Participants were reimbursed pro-rata at £7.50/hour (¥180 total for survey in Japan). An attention check was included at the beginning of the survey; those who failed were excluded from analysis. In the UK survey, 190 participants took part, of whom 186 passed the attention check and provided full data were included in the analysis. In the Japanese survey, 360 participated (346 valid responses were included in analysis). All materials and the complete data can be found at: https://osf.io/kv6nu/. Regarding demographic information, In the UK sample (N = 186), 27% were male and 64% were female, while 9% were unknown. 47% were in the range of age 20–40 and 42% were aged between 40–49 or over 50, while 11% were unknown. In the Japanese sample (N = 346). 43% were male and 49% were female, while 9% were unknown. 25% were in the range of age 18–40 and 27% were aged between 40–49 or over 50, while 48% were unknown (S1 File). Demographic information was collected from the survey providers, and not linked to individual responses.

2.2 Purchase Preference: Personal Ethical Setting

The questionnaire was created in English and translated into Japanese (available at: osf.io/kv6nu) (For survey flow see S4 Text in S1 File). At the beginning, we provided background information: participants were informed that driverless cars were estimated to reduce traffic accidents by 90–94% but would need to be programmed in advance how to respond to any collisions. For the purposes of the survey, they were asked to imagine that in ten years' time, driverless cars would have the same cost and same features as a regular car but would be safer. They were asked to ignore any concerns about data privacy or liability. Their initial perceived likelihood of purchasing a driverless car (on a Likert scale from 1—Not likely to 7—Extremely likely) was assessed as a baseline measure.

Subsequently, participants were given information about programming options for driverless cars' responses to collisions. Three programming algorithms were described: 'Save the Pedestrians' [always save pedestrians in a collision between driverless car and pedestrian]; 'Save the Occupants' [always save occupants of the car]; and, 'Save the Most' [always save the greater number of people]. We included edited images from the Moral Machine study [14] to help participants understand the different algorithms. The survey then asked them which of these programming models they would prefer if they were to purchase such a car. They were given the three programming algorithms (named, and briefly described), as well as the option of 'Random choice', wherein the driverless car would randomly choose to save either occupants or pedestrians.

2.3 Framing Effects and Purchase Preference

Next, they were asked to imagine (and to indicate their purchasing preference), first if they were buying a car in a situation in which they had a young family member who would often be a passenger, and second in a situation in which they had a young family member who would often be a pedestrian. As before, they were given the option of the three driverless car algorithms ('Save the Pedestrians', 'Save the Occupants', 'Save the Most') or 'Random Choice'.

2.4 Mandatory Ethical Setting: Purchase and Moral Algorithm Preference

In the next section, participants were asked to indicate how likely (on a scale from 1 (Not very likely) to 7 (extremely likely)), they were to buy a driverless car if all such cars were programmed with the same algorithm (i.e., mandatory ethical setting). For example, they were told that all cars were programmed to 'Save the Most'. This question was repeated to assess likelihood of purchase for each of the three algorithms ('Save the Pedestrians',' Save the Occupants', 'Save the Most').

They were lastly asked which programming model they believed should be adopted if all driverless cars were going to be programmed in the same way (their Moral Algorithm Preference), (they were given the

three algorithms as well as the option of 'Random Choice'). They were reminded that they might be either occupants or pedestrians.

2.5 Price Sensitivity

Finally, a discrete choice experiment was conducted. Participants were presented with a series of choices between two driverless cars that feature different programming models but have different price tags and asked which car they would buy. It has been estimated that most new cars in the UK sell from about £12,000 to £23,000 [30]. We used that price range, starting with £15,000. There were 36 combinations for the price tests due to the three different models and three different prices (£15,000, £19,000 and £23,000; ¥2,250,000, ¥2,850,000, and ¥3,450,000). Fig 8.2 shows an example question:

Fig 8.2 An example question for the discrete choice experiment evaluating price sensitivity and preference for driverless car algorithms.

https://doi.org/10.1371/journal.pone.0275812.g001

2.6 Model Description—Willingness to Pay for the Purchase

In our model, the utility difference between Car A and Car B can be described as the difference in price and the difference in safety programming:

$$\Delta U(A, B) = (Price_B - Price_A) + (Prog_A - Prog_B) \quad (1)$$

Where $Price_A$ is the price of Car A (e.g., £15000) and $Prog_A$ is the utility associated with the programming of Car A (e.g., Utilitarian).

There are three programming options: 'Save the Occupants', 'Save the Pedestrians', 'Save the Most' (utilitarian), resulting in two independent parameters. We set the utility of the 'Save the Most' option to zero, making it the reference point for the other two programming options. And we expressed the preferences for each of these programming options ('Save the Occupants' = Occ; 'Save the Pedestrians' = Ped) in units of thousands of pounds.

How did we turn a utility difference $\Delta U(A, B)$ into a choice probability? Roughly, if the utility difference is large and positive, participants ought to choose A with probability ~1. And if the utility difference is large and

negative, participants ought to choose B with probability ~1. To model the choice probability, we scaled the utility by multiplying by a parameter $\sqrt{\tau}$, and then applied an inverse probit transformation. This allowed us to go from the utility difference (Eq 1) to the probability of choosing A given options A and B,

$$p_{AB}(A) = probit^{-1}(\sqrt{\tau}\,\Delta U(A, B)) = \int_{-\infty}^{\Delta U(A,B)} \sqrt{\frac{\tau}{2\pi}} e^{\frac{x^2}{2}} dx \tag{2}$$

The parameter τ represents how sensitive participants were to changes in utility, so how quickly choices switched between definitely A and definitely B as the utility varied. Lower values mean participants spent more time in the 'unsure' region.

At the individual level the choices are binary, so our choice probability that came from the model resulted in just a single yes/no response. We can model this by assuming the data comes from a single trial where the choice probability is given by $p_{AB}(A)$, in other words,

$$p(Data|\Delta U) \sim Bernouli(p_{AB}(A)) \tag{3}$$

To model this in a hierarchical way means a) every individual has their own values for the parameters *Ped, Occ,* τ, b) These individual values can be assumed to be drawn from some population level distributions. Each of these population level distributions has an associated mean, *HPed, HOcc,* $H\tau$, and variance. More precisely,

$$Ped(i) \sim \phi(HPed, Ht), \quad Occ(i) \sim \phi(HOcc, Ht), \quad \tau(i) \sim \phi(H\tau, Ht_2) \tag{4}$$

Where $\phi(\mu, \tau)$ is the pdf of the normal distribution with mean μ and variance $1/\tau$. The idea here is that we allow for individual variation in utilities and sensitivity, but we constrain these differences so that information about the preferences of one participant is still weakly informative about the preferences of the others. *HPed, HOcc,* $H\tau$*, Ht, and* Ht_2 are population level parameters (hyperparameters), which govern the distribution of utilities and sensitivity in the population. These are the things to be determined by the model fitting, and we assign them the following priors,

$$HPed \sim \phi\left(0, \frac{1}{10}\right), HOcc \sim \phi\left(0, \frac{1}{10}\right), H\tau \sim \gamma\left(\frac{1}{2}, \frac{1}{2}\right), Ht \sim \gamma\left(\frac{1}{2}, \frac{1}{2}\right), Ht_2 \sim \gamma\left(\frac{1}{2}, \frac{1}{2}\right). \tag{5}$$

where $\gamma(a, b)$ is the pdf of the gamma distribution with shape parameters, a, and b.

2.7 Model Implementation

The model was fit to the data via Bayesian methods using JAGS (Plummer, 2003 [31]), using a form of Markov Chain Monte Carlo (MCMC) [32]. Fits used three MCMC chains and 50000 MCMC samples, with a burn in of 5000 samples. Chain convergence was assessed using the $\hat{R}$ statistic, and all chains had good convergence by this metric. We report means and HDIs for the posteriors of the hyperparameters, and distributions of the Ped and Occ parameters across participants and trials.

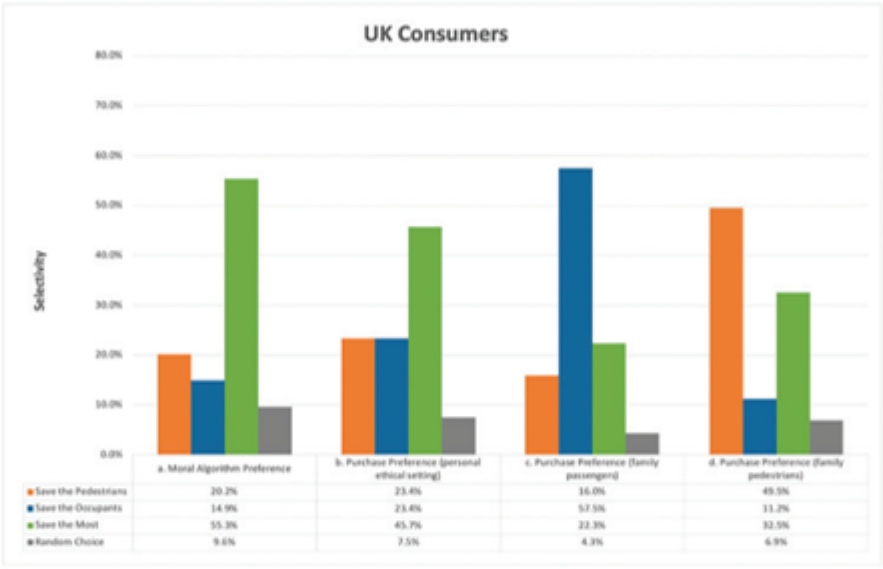
3. Results

Statistical analysis was conducted using SPSS, Version 28, and computational model analysis were performed with Matlab.

3.1 Algorithm and Purchase Preference–Which Algorithm Should Cars have?

The largest group of participants indicated the same response when asked which algorithm they believed should be programmed into all driverless cars (Moral Algorithm Preference), and which algorithm they would prefer to purchase if able to choose (purchasing preference; personal ethical setting) in both countries

(Fig 8.3). In the UK, the largest group selected 'Save the Most' (55.3% answered 'Save the Most' should be programmed and 45.7% answered it was their preferred algorithm), while in Japan, the largest group selected "Save the Pedestrians' (54.6% and 56.9% respectively). A minority of participants wished to purchase the other models (UK: 'Save the Occupants' 23.4%, 'Save the Pedestrians', 23.4%; Japan: 'Save the Occupants' 15.6%, 'Save the Most' 24.6%).



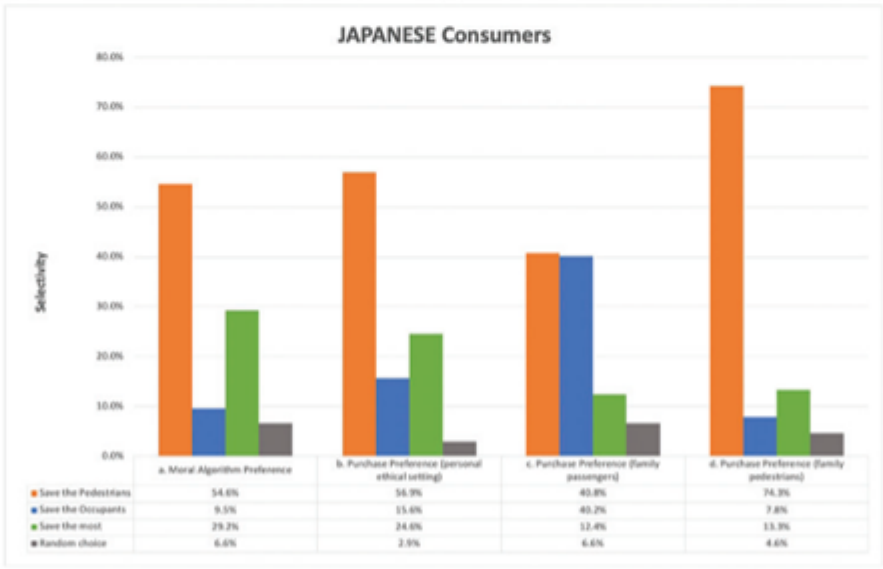


Fig 8.3 Participant preference among UK and Japanese participants for driverless car algorithms when asked: a. which algorithm should be programmed (if all cars programmed identically) b. which they would personally prefer to purchase if able to choose c. Their purchase preference if they imagined having young family who would often be passengers d. Their purchase preference if they imagined having young family who would often be pedestrians.

https://doi.org/10.1371/journal.pone.0275812.g002

While Moral Algorithm Preferences aligned with purchasing preference in each country, there was a statistically significant difference between Purchase Preferences and Moral Algorithm Preference: Moral Algorithm Preference for the UK (McNemar-Bowker (6 N = 186) = 17.60, p = .007) and for Japan (McNemar-Bowker (6) = 23.655, p = .001, N = 346).

3.2 The Effect of Framing on Purchase Preference

Next, we examined the effects of different question framing. When asked to imagine that one's family would often be in the car, a higher proportion of respondents in both countries selected to purchase 'Save the Occupants' compared to when the question was presented without such a frame (UK: 57%, McNemar-Bowker (6, N = 186) = 68.415, p < .0001; Japan: 40.2%, McNemar-Bowker (6, N = 346) = 103.814, p < .001). Similarly, there was a significant increase in preference for 'Save the Pedestrians' when asked to imagine one's family often being pedestrians compared to when the question was presented without such a frame (UK: 50%, McNemar-Bowker (6, N = 186) = 50.509, p < .0001; Japan: 74.3% McNemar-Bowker (6, N = 346) = 63.800, p < .001). Responses to both framings were different between the UK and Japan (p < .001).

3.3 The Influence of Mandatory Ethical Settings on Willingness to Purchase

As a baseline question, we asked participants how likely it is that they would purchase a driverless car in ten years' time (if their current car needed replacing) without specifying any safety programming. In the UK, the median response on the 1 (not at all likely) to extremely likely (7) scale was 5, with a Mean of 5.13 (SE = .119). In Japan, the median response was 6, with a Mean of 5.66 (SE = .068). Overall, 74% of the UK participants, and 82% of the Japanese indicated that they were likely (response >4) to buy a driverless car (neither likely nor unlikely: UK 10%, Japan 9%). Japanese participants had a higher willingness to purchase a driverless car than UK participants, t(530) = 4.166, p < .001.

Next, we examined how consumers' willingness to purchase driverless cars changes when a single algorithm is programmed for all driverless cars. For all three mandated ethical settings, there was reduced willingness to purchase compared to the baseline question indicating that some participants would only purchase if their preferred programming algorithm were available (when 'Save the Most' algorithm is only available—UK participants (M = 4.14, SE = .128, t(185) = 7.918, p < .001) and Japanese participants (M = 4.17, SE = .07, t(345) = 16.69, p < .001); when 'Save the Pedestrians' algorithm is only available—UK participants (M = 3.58, SE = .1249, t(185) = 10.925, p < .001) and Japanese participants (M = 4.77, SE = .07, t(345) = 10.72, p < .001); when 'Save the Occupants' algorithm is only available—UK (M = 4.134, SE = .123, t(185) = 7.59, p < .001) and Japanese participants (M = 4.12, SE = .093, t(345) = 17.505, p < .001)).

If a mandatory ethical setting were adopted, 47% of UK respondents would buy a car programmed to save occupants (a drop from 74% baseline) (Fig 8.4). 60% of Japanese respondents would purchase a car programmed to save pedestrians (a drop from 82% baseline non-specified). Overall, participants' willingness of the purchase was affected by the type of algorithm in both countries.

By tracking each participant's purchasing preference on the three algorithms, a Venn diagram was created to illustrate the distribution of consumers' preferences for those three algorithms (Fig 8.5). A small proportion of participants were likely to buy a driverless car whichever algorithm was available (UK 11.8%, Japan 13.6%), and a minority were not willing to purchase regardless of algorithm (UK 23.1%, Japan 21.1%). In the UK, two thirds of respondents (67.8%) were willing to purchase either 'Save the Occupants' or 'Save the Most' models. In Japan, 72.3% were willing to purchase either 'Save Pedestrians' or 'Save the Occupants'.

https://doi.org/10.1371/journal.pone.0275812

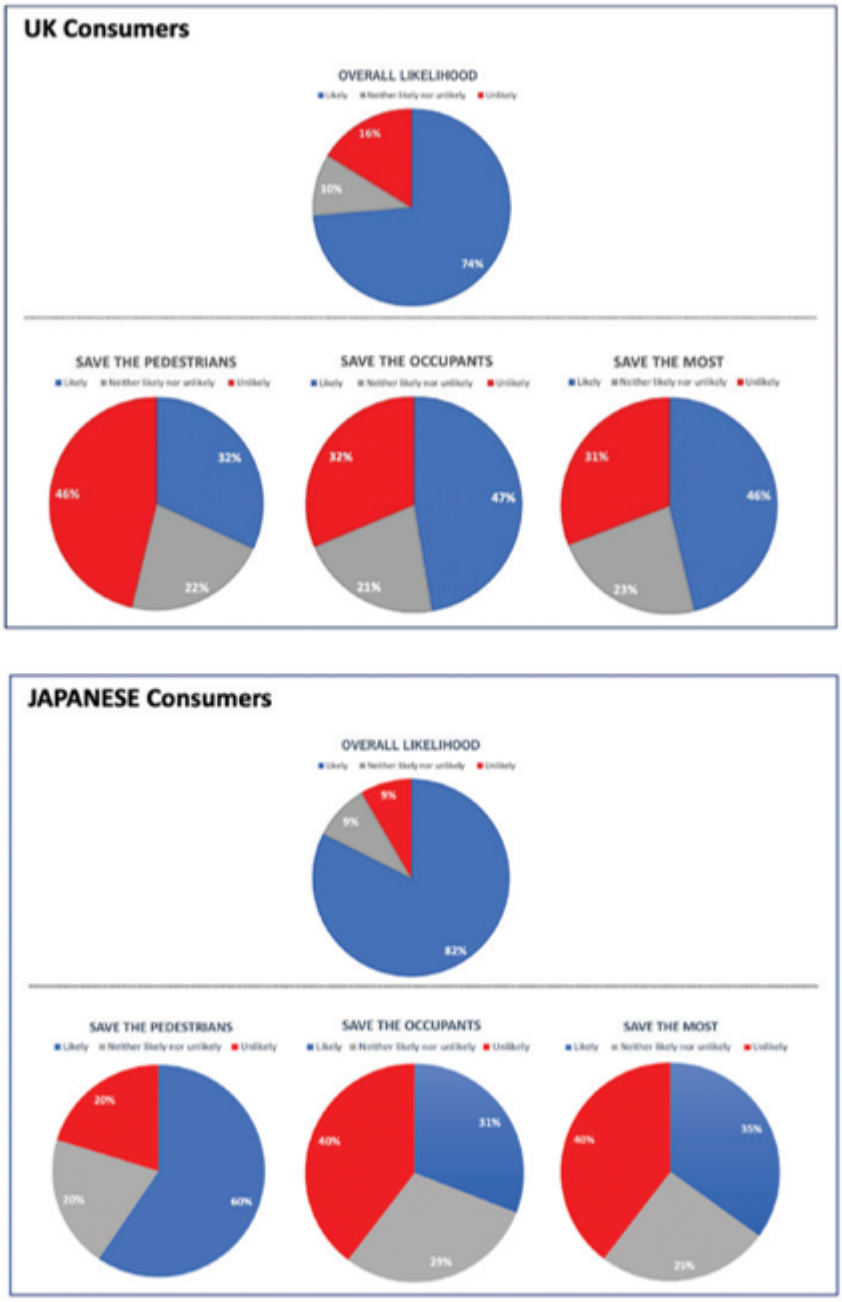


Fig 8.4 The overall likelihood of purchasing a driverless car and the likelihood of purchasing a driverless car if a mandatory ethical setting were used and particular algorithms were the only option available.

https://doi.org/10.1371/journal.pone.0275812.g003

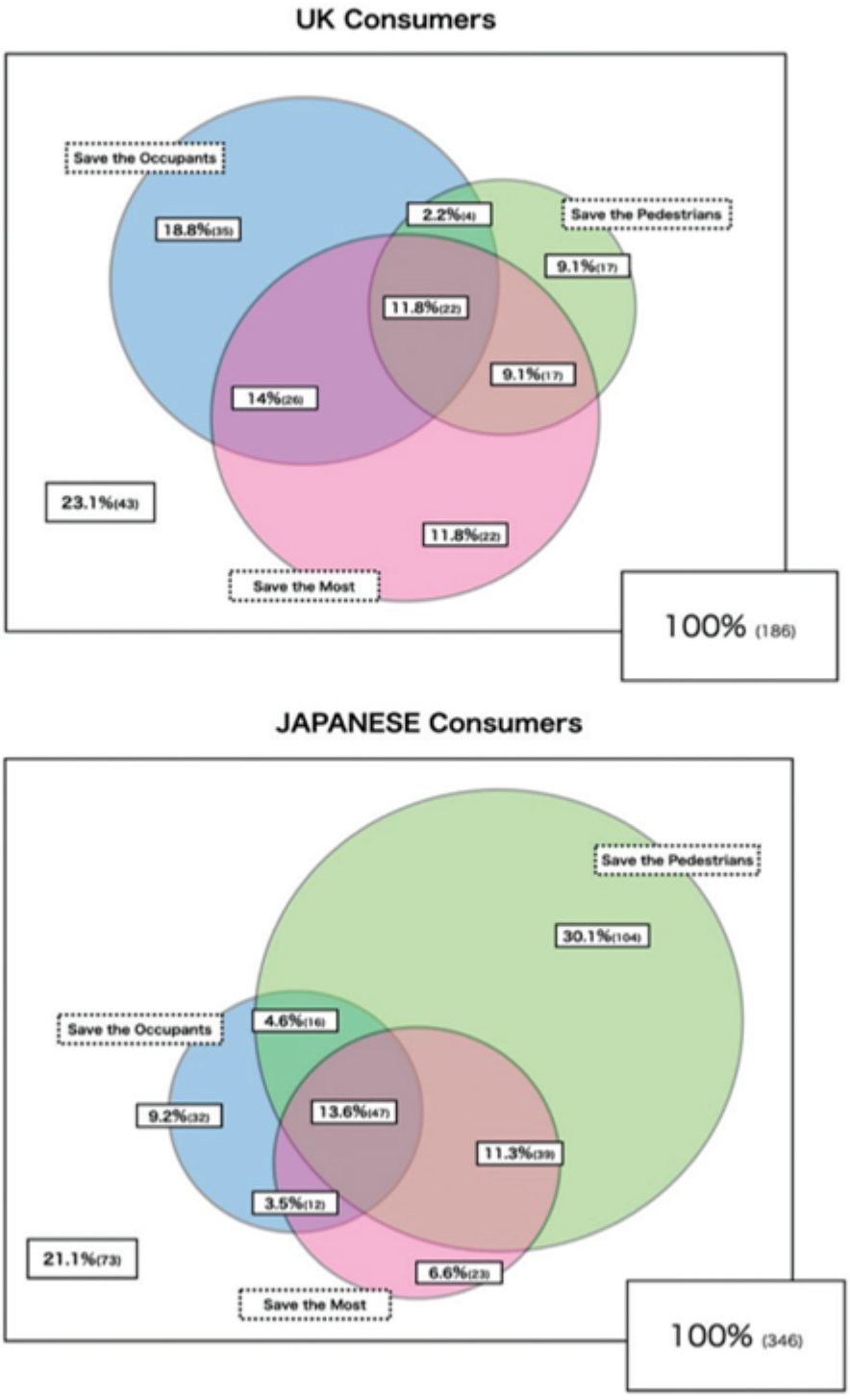


Fig 8.5 Allocation of consumers' preferences for three different algorithms. Respondents who were "unlikely" or "neither likely/unlikely" to purchase any model are indicated outside the Venn diagram (the figures in parentheses refer to the actual number of participants).

https://doi.org/10.1371/journal.pone.0275812.g004

3.4 Willingness to Pay: Purchase Preferences

The model was able to predict well which decisions UK and Japanese participants would make on each of the 36 scenarios (see S1 File). Examining the estimates for the population level parameters for the preference strength, we observed that for the UK population, going from a utilitarian program to one favouring pedestrians was equivalent to a price increase of £10,700, while going from a utilitarian option to one favouring occupants was equivalent to a price increase of around £3,170 (Fig 8.6). A different picture emerged for the Japanese population. Here, the model estimated that going from a utilitarian program to one favouring pedestrians was equivalent to a price decrease of £4,180, while going from a utilitarian option to one favouring occupants was equivalent to a price increase of around £6,470.

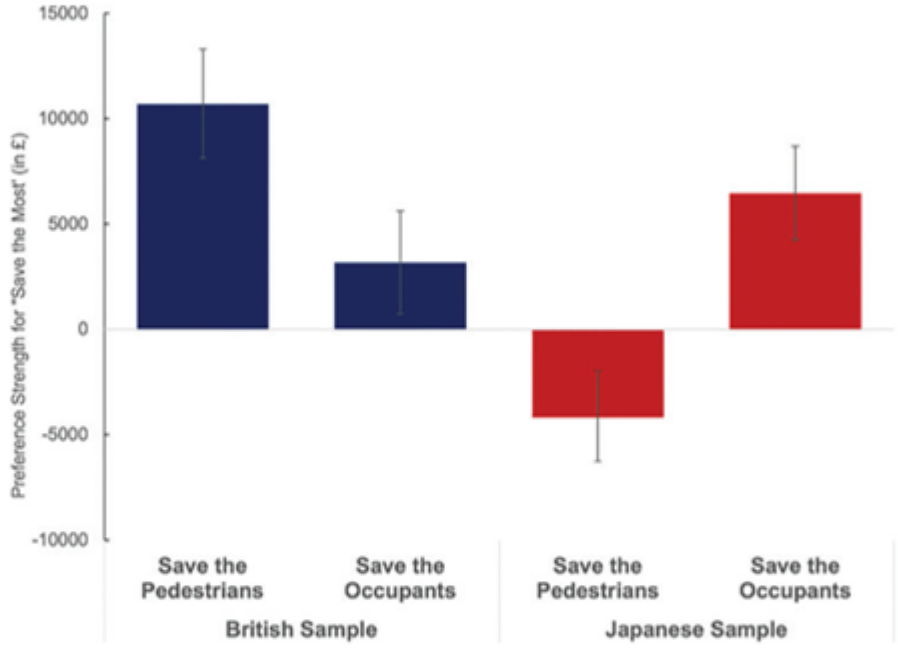


Fig 8.6 Average preference strength of 'Save the Pedestrians' and 'Save the Occupants' over 'Save the Most'. (A positive price changes means that participants preferred 'Save the Most', and would require a price discount to choose the alternative).

https://doi.org/10.1371/journal.pone.0275812.g005

Comparing the population estimates (S2 Table in S1 File), the largest difference between the UK and the Japanese population was the utility assigned to 'Save the Pedestrians' programming. For the UK sample this was assigned much lower utility than the 'Save the Most' or 'Save the Occupants' options. For the Japanese sample in contrast, 'Save the Pedestrians' was significantly more attractive than either 'Save the Occupants' or 'Save the Most' options. There was also a significant difference in the utility assigned to the 'Save the Occupants' programming, with the Japanese sample rating this as significantly worse than the UK sample. Finally, we also found that Japanese participants showed a greater consistency in their utility judgments (t parameter: Japan .0045 versus UK .0034) significantly higher, and a less noisy decision-making process (Tau parameter: Japan: .388 versus UK.291), indicating that Japanese preferences were more stable than the UK ones.

In each country, large individual differences existed in the UK and the Japanese Sample (S2 Fig in S1 File). When we examined the distribution of preferences values within our UK sample for 'Save the Pedestrians'

over the 'Save the Most', we can see that a minority of 25 participants preferred a 'Save the Pedestrian' programming to a 'Save the Most' programming. Twenty-five participants would have paid up to £40,000 pounds more to get a car with 'Save the Most', and avoid a car with 'Save the Pedestrian' programming (S2 Fig in S1 File). In the Japanese sample, 149 participants preferred the 'Save the Most' over the 'Save the Pedestrian' safety programming. Here, 42 participants would have paid up to £10,000 to get a 'Save the Pedestrian' over a 'Save the Most' programming, and 10 participants up to £20,000 more. The most popular safety programming in each country was not endorse by everyone, and mandatory settings might deter large minorities from purchasing driverless cars. Even more pronounced is this effect when we examine the distribution of preference values in both samples for 'Save the Occupants' over 'Save the Most'. Here, we can see that in the UK sample, 65 participants preferred such programming over 'Save the Most', and in the Japanese sample, 105 participants preferred 'Save the Occupants' over 'Save the Most'.

4. Discussion

In this international online survey, we found striking differences between potential consumers in the UK and Japan in their preferences for the programming of driverless cars and their hypothetical purchase. A majority of UK participants (55.3%) supported the programming of driverless cars to 'Save the Most' lives in a collision. In contrast, a majority of Japanese respondents (54.6%) supported algorithms that would prioritise the saving of pedestrians. Purchasing preferences in both countries were highly sensitive to framing with shifts (in favour of saving family members) when asked to imagine that their family members would often be either occupants or pedestrians. Importantly, in both countries, willingness to purchase was lower when ethical settings were mandated to a single algorithm, compared to baseline willingness to purchase and compared to the proportion willing to purchase when given options of driverless car algorithms. Respondents from the UK were divided (a similar proportion being willing to buy a car programmed to 'Save the Most' and to 'Save the Occupants'), while a majority in Japan were willing to purchase a driverless car that would prioritise pedestrians. In both countries, price sensitivity tracked general preferences, and participants generally placed a significant price premium on their personally preferred algorithm.

4.1 Cultural Variation in Values Applied to Driverless Cars

The overall Moral Algorithm Preference identified in our study is consistent with previous studies. Surveys (largely in European and North American populations) have generally supported driverless algorithms that would save the most lives (Table 8.8), and this was described as a globally shared preference in the MME [14]. We observed a similar pattern in the responses from our UK respondents. However, within the MME, there was a significant difference between respondents from a 'Western' cluster of countries and those from an 'Eastern' cluster [14]. Our finding that Japanese respondents had a much stronger preference for prioritising pedestrians is consistent with that prior global survey. It may reflect communitarian values in Japan and strong senses of social responsibility and conformity [33]. It is also possible that the Japanese preference reflects recent media attention and social concern about accidents involving elderly drivers and pedestrians [34], or lower levels of vehicle ownership in Japan. Further studies are needed to know the reasons for Japanese purchasing behaviours.

4.2 Algorithm Versus Purchase Preferences

We found a difference between Moral Algorithm Preferences (which algorithm they think should be programmed) and purchasing preferences. However, such purchasing preferences (in the absence of family framing) generally tracked participants' views about what should be programmed overall (Fig 2). This is somewhat in contrast to Bonnefon et al., who found a marked fall in support for a utilitarian algorithm when participants were asking about purchase [19]. However, we found that Purchase Preferences were highly sensitive to question framing. Past studies asking about general Moral Algorithm Preference have

found some effect of framing perspective on responses (i.e., asked to imagine or respond as pedestrian vs occupant of car), but generally preserved utilitarian responses. We found much stronger preferences when participants were asked to imagine family members involved in accidents. It may be that our questions about Purchase Preferences are more individually directed, and therefore more sensitive to personal circumstances. Alternatively, it may be that the wording of questions in our survey primed respondents to alter their responses (possibly in a perceived socially desirable direction).

4.3 Mandatory Ethical Setting

We were interested to explore the impact of a mandatory ethical setting on overall willingness to purchase driverless cars. In Bonnefon et al.'s study, US consumers indicated aversion to governmental regulation on what algorithm to be programmed into driverless cars, which was pointed out as a concern for mandatory ethical setting. Although we did not ask the UK and Japanese consumers' attitude towards governmental regulation, our data suggest that mandatory ethical settings discourage consumer uptake since for many consumers', their decisions whether or not to buy driverless cars were contingent on which accident algorithm is available. For example, from our UK data, a mandatory ethical setting with any type of algorithm reduced by 20% the number of consumers willing to buy driverless cars. It appeared that the largest number of respondents would buy a driverless car if given the option of choosing between the three different algorithms, or what we call a personal ethical setting. Very few respondents indicated a preference for abdicating responsibility by having the car randomly choose to save either pedestrians or occupants.

4.4 The Utility Paradox and the Ethics of Personal Ethical Settings

Overall, our results appear to support what we have labelled the 'utility paradox' in driverless car algorithms. In our sample, across two countries, algorithms designed to 'Save the Most' lives would actually lead to lower uptake of driverless cars and thus save fewer lives in practice, given the ability of driverless cars to reduce trade-off accidents in total. In Japan, that was because most respondents preferred driverless cars to prioritise pedestrians. In the UK, respondents were divided as to which algorithm they would prefer. Any mandated ethical setting attracted a smaller proportion of hypothetical consumers than those who indicated (without an algorithm specified) they were willing to purchase a driverless car.

This finding suggests that making driverless cars available with a personal ethical setting may yield a higher uptake and consequently have the greatest overall impact on road casualties, at least in the first instance. This is a libertarian option [35], but would also be predicted to have the greatest utility (and therefore be supported by utilitarianism). As some of the authors have argued elsewhere, from behind a veil of ignorance, it is rational to select the policy which saves the most lives (including those of the individual and family members) [36]. Thus, a broad contractualist approach would also support making driverless cars available with a personal ethical setting when they are introduced into general circulation. From this point of view, a personal ethical setting has the advantage that while it would be supported by utilitarians and contractualists by maximising driverless car uptake, it would also be aligned with liberalism by providing individual value-based choices.

This argument in favour of a personal ethical setting is dependent on purchasing choices, and our survey suggests that such choices may be susceptible to context and framing effects. It may be more of an advantage for some countries than others (in our Japanese participants, a 'Save the Pedestrians' mandatory setting appeared to have high support). One way for governments to increase uptake may be to subsidise the cost of driverless cars in general, or of particular algorithms (for example, in a way similar to subsidies for electric cars). For example, our willingness to pay analysis suggests that in Japan, a price subsidy of £4000 might be required to shift preferences to a 'Save the Most' algorithm. Another possibility (which would merit further study) may be to adapt algorithms to include a combination of values (for example, prioritising car occupants if passengers in the car, and similar numbers to pedestrians, while otherwise aiming to 'Save the Most'). Yet another option

https://doi.org/10.1371/journal.pone.0275812

would be to introduce driverless cars with a personal ethical setting in the initial stage, then to shift to a (e.g. 'Save the Most') mandated ethical setting once non-automated cars have been largely or entirely replaced.

There are other ethical considerations for a personal ethical setting. One concern may be that use of such an option would lead owners of cars to be morally responsible (and potentially liable for damages) in the event of collision fatalities, perhaps analogously to driving beyond the speed limit [37]. It may be that insurance costs would be higher for those who do not choose to purchase 'Save the Most' algorithms [37]. Allowing individuals to choose their ethical setting might be socially divisive or lead to stigma. On the other hand, a personal ethical setting allows consumers to exercise their freedom and autonomy. Given a range of different plausible ethical responses (particularly where there is not a large difference in the number of people under threat), a libertarian response [38] might allow individuals to incorporate their personal values (within limits) into their vehicle programming. It is potentially important to allow individuals to take moral responsibility for their actions when using artificial intelligence. The Japanese Cabinet Secretariat accepted this concept and pointed out that "when using AI, people must judge and decide for themselves how to use it" [39]. That would allow individuals to choose to sacrifice their own (and their passengers) to save the greatest number of lives in a collision. Individual autonomy arguably should be respected and is especially crucial in a situation where the person's life is at stake.

One objection is that autonomy can be curtailed in the public interest or for the sake of public health. Yet our research indicates a happy convergence of personal autonomy and public interest in the UK and Japan. Based on plausible assumptions about the high-safety level of driverless car system, the rarity of trade-off accidents and contingent on purchasing behaviors, the most lives would be saved if people were allowed to express their autonomy (unless governments banned non-autonomous vehicles and mandated an algorithm of 'Save the Most Lives').

4.5 Limitations and Future Directions

There are some limitations to the generalisability of these results. Online surveys have inherent selection biases. Although we used participants pools that have previously been validated in behavioural research, those who completed our surveys may not be representative of the wider population. (On the other hand, a younger, internet-using audience may overlap with the market for driverless cars). While the sample size was similar to many previous studies in this area (Table 8.8), a larger sample would have allowed more confident estimates of population attitudes. Only those who passed an attention check were included in the analysis, but, some of those included in the analysis may not have given careful thought to each question. This might be a problem especially in the discrete choice experiment.

As we found in our survey, respondents may be susceptible to the framing of the questions, and the specific wording of our study may have influenced results. (However, the fact that the global preferences among UK and Japanese respondents were consistent with those seen in the MME [14] and other related studies [40], supports the validity of responses). We were interested in the impact of driverless car algorithms on purchase preferences. However, stated intentions may not represent actual purchase choices. Participants may change their mind in the future or fail to imagine what they will actually do in the real situation.

5. Conclusions

Fully automated driverless cars are not yet commercially available. But the results of this study provide some insights that may be important for their regulation and development. Public views about the safety and ethics of driverless cars will be crucial for their acceptance. It may be that how such cars are programmed to respond may need to differ between countries–depending on the prevailing values of the community. It may also be that permitting some ethical choice in driverless car programming would be valuable.

https://doi.org/10.1371/journal.pone.0275812

The programming of driverless cars is fundamentally an ethical decision. We must take responsibility for the goals of such programming, whether it is saving the greatest number, or considering other factors such as age, responsibility, etc [40]. Some kind of democratic process preferably informed by ethical procedures [41] is necessary to arrive at how lives should be valued. Regulation and mandating of programming and purchase of driverless vehicles to achieve these ethical goals (once they have been derived) may be justifiable. However, in the absence of mandated replacement of non-automated vehicles, it is essential to explore the psychosocial effects of policy on purchasing behaviour because the most important factor affecting the well-being and autonomy of road users and those affected by road use is likely to be uptake of driverless cars. If we were to minimise casualties overall, it might not be wise to adopt a 'Save the Most' algorithm as a mandated ethical setting because of the potential to decrease driverless car uptake (the utility paradox). We have argued that allowing freedom of choice of ethical setting in driverless cars would potentially be supported by utilitarian, contractualist and libertarian ethical theories. It may be the most ethical policy.

Supporting Information

S1 File. It consists of all the supporting files for this manuscript.
(DOCX)

Acknowledgments

We would like to show our gratitude to Professor Tony Hope (an emeritus fellow of St Cross college at University of Oxford) for sharing his wisdom with us. Our discussion about connecting philosophical debates and empirical research was enriched by his critical insights.

Author Contributions

Conceptualization: Kazuya Takaguchi.
Data curation: Andreas Kappes.
Formal analysis: Kazuya Takaguchi, Andreas Kappes, James M. Yearsley.
Funding acquisition: Julian Savulescu.
Investigation: Kazuya Takaguchi, Tsutomu Sawai.
Methodology: Andreas Kappes.
Project administration: Dominic J. C. Wilkinson.
Resources: Tsutomu Sawai.
Software: Andreas Kappes.
Supervision: Dominic J. C. Wilkinson, Julian Savulescu.
Validation: Dominic J. C. Wilkinson.
Visualization: Kazuya Takaguchi.
Writing – original draft: Dominic J. C. Wilkinson.
Writing – review & editing: Kazuya Takaguchi, Dominic J. C. Wilkinson, Julian Savulescu.

References

1. Benson AJ, Tefft BC, Svancara AM, Horrey WJ. Potential Reductions in Crashes, Injuries, and Deaths from Large-Scale Deployment of Advanced Driver Assistance Systems. Res Brief. 2018 [cited 29 Aug 2022]. Available: https://trid.trb.org/view/1566022
2. WHO. Global status report on road safety 2018. 2018 [cited 27 Apr 2022]. Available: https://www.who.int/publications-detail-redirect/9789241565684

3. Verkehrsunfälle—Fachserie 8 Reihe 7–2017. In: Statistisches Bundesamt [Internet]. [cited 27 Apr 2022]. Available: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Verkehrsunfaelle/Publikationen/Downloads-Verkehrsunfaelle/verkehrsunfaelle-jahr-2080700177004.html
4. Department for Transport. Contributory Factors for Reported Road Accidents (RAS50). 2013. Available: https://www.gov.uk/government/statistical-data-sets/ras50-contributory-factors
5. Singh S. Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey (Traffic Safety Facts Crash·Stats. Report No. DOT HS 812 115). 2015. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115.pdf
6. Utriainen R. The potential impacts of automated vehicles on pedestrian safety in a four-season country. J Intell Transp Syst. 2021;25: 188–196. https://doi.org/10.1080/15472450.2020.1845671
7. Combs TS, Sandt LS, Clamann MP, McDonald NC. Automated Vehicles and Pedestrian Safety: Exploring the Promise and Limits of Pedestrian Detection. Am J Prev Med. 2019;56: 1–7. https://doi.org/10.1016/j.amepre.2018.06.024 PMID:30337236
8. LaFrance A. Self-Driving Cars Could Save 300,000 Lives Per Decade in America. In: The Atlantic [Internet]. 29 Sep 2015 [cited 27 Apr 2022]. Available: https://www.theatlantic.com/technology/archive/2015/09/self-driving-cars-could-save-300000-lives-per-decade-in-america/407956/
9. Gogoll J, Müller JF. Autonomous Cars: In Favor of a Mandatory Ethics Setting. Sci Eng Ethics. 2017;23: 681–700. https://doi.org/10.1007/s11948-016-9806-x PMID:27417644
10. Mathilde C. Average cost of road casualties and accidents by severity in Great Britain in 2020. 2020. Available: https://www.statista.com/statistics/322862/average-cost-of-road-accidents-and-casualties-in-great-britain-uk/
11. Healey J, Mays K. Which cars park themselves best? Challenge results. In: USA Today [Internet]. 2012 [cited 27 Apr 2022]. Available: https://www.usatoday.com/story/money/cars/2012/12/06/self-parking-cars-challenge/1743199/
12. Foot F. The Problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices. Oxf Rev 5. 1967; 5–15.
13. BMVI Ethical Commission. Automated and Connected Driving. 2017. Available: https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission-automated-and-connected-driving.pdf?__blob=publicationFile
14. Awad E, Dsouza S, Kim R, et al. The Moral Machine experiment. Nature. 2018 Oct;563: 59–64.
15. Bigman YE, Gray K. Life and death decisions of autonomous vehicles. Nature. 2020;579: E1–E2. https://doi.org/10.1038/s41586-020-1987-4 PMID:32132695
16. Frank D-A, Chrysochou P, Mitkidis P, Ariely D. Human decision-making biases in the moral dilemmas of autonomous vehicles. Sci Rep. 2019;9: 13080. https://doi.org/10.1038/s41598-019-49411-7 PMID:31511560
17. Kallioinen N, Pershina M, Zeiser J, Nosrat Nezami F, Pipa G, Stephan A, et al. Moral Judgements on the Actions of Self-Driving Cars and Human Drivers in Dilemma Situations From Different Perspectives. Front Psychol. 2019;10: 2415. https://doi.org/10.3389/fpsyg.2019.02415 PMID:31749736
18. Mayer MM, Bell R, Buchner A. Self-protective and self-sacrificing preferences of pedestrians and passengers in moral dilemmas involving autonomous vehicles. PLOS ONE. 2021;16: e0261673. https://doi.org/10.1371/journal.pone.0261673 PMID:34941936

19. Bonnefon J-F, Shariff A, Rahwan I. The Social Dilemma of Autonomous Vehicles. Science. 2016;352. https://doi.org/10.1126/science.aaf2654 PMID:27339987

20. Liu P, Liu J. Selfish or Utilitarian Automated Vehicles? Deontological Evaluation and Public Acceptance. Int J Human–Computer Interact. 2021;37: 1231–1242. https://doi.org/10.1080/10447318.2021.1876357

21. Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, et al. Reply to: Life and death decisions of autonomous vehicles. Nature. 2020;579: E3–E5. https://doi.org/10.1038/s41586-020-1988-3 PMID:32132696

22. Bergmann LT, Schlicht L, Meixner C, König P, Pipa G, Boshammer S, et al. Autonomous Vehicles Require Socio-Political Acceptance—An Empirical and Philosophical Perspective on the Problem of Moral Decision Making. Front Behav Neurosci. 2018;12: 31. https://doi.org/10.3389/fnbeh.2018.00031 PMID:29541023

23. Faulhaber AK, Dittmer A, Blind F, Wächter MA, Timm S, Sütfeld LR, et al. Human Decisions in Moral Dilemmas are Largely Described by Utilitarianism: Virtual Car Driving Study Provides Guidelines for Autonomous Driving Vehicles. Sci Eng Ethics. 2019;25: 399–418. https://doi.org/10.1007/s11948-018-0020-x PMID:29357047

24. Li S, Zhang J, Li P, Wang Y, Wang Q. Influencing Factors of Driving Decision-Making Under the Moral Dilemma. IEEE Access. 2019;7: 104132–104142. https://doi.org/10.1109/ACCESS.2019.2932043

25. Pugnetti C, Schläpfer R. Customer Preferences and Implicit Tradeoffs in Accident Scenarios for Self-Driving Vehicle Algorithms. Risk Financ Manag. 11(2): 2018.

26. Wintersberger P, Prison A-K, Riener A, Hasirlioglu S. The experience of ethics: Evaluation of self harm risks in automated vehicles. 2017 IEEE Intelligent Vehicles Symposium (IV). 2017. pp. 385–391. https://doi.org/10.1109/IVS.2017.7995749

27. Chandler J, Shapiro D. Conducting Clinical Research Using Crowdsourced Convenience Samples. Annu Rev Clin Psychol. 2016;12: 53–81. https://doi.org/10.1146/annurev-clinpsy-021815-093623 PMID:26772208

28. Faul F, Erdfelder E, Buchner A, Lang A-G. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behav Res Methods. 2009;41: 1149–1160. https://doi.org/10.3758/BRM.41.4.1149 PMID:19897823

29. Majima Y. The Feasibility of a Japanese Crowdsourcing Service for Experimental Research in Psychology. SAGE Open. 2017;7: 2158244017698731. https://doi.org/10.1177/2158244017698731

30. Yurday E. Average Cost of Cars UK 2021. 2021 [cited 3 Sep 2022]. Available: https://www.nimblefins.co.uk/cheap-car-insurance/average-cost-cars-uk

31. Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna, 20–22 March 2003, 1–10.

32. van Ravenzwaaij D, Cassey P, Brown SD. A simple introduction to Markov Chain Monte-Carlo sampling. Psychon Bull Rev. 2018;25: 143–154. https://doi.org/10.3758/s13423-016-1015-8 PMID:26968853

33. Inoue T. The Poverty of Rights-Blind Communality: Looking Through the Window of Japan. BYU Law Rev. 1993;1993: 517–551.

34. Shimada T. Elderly driver in fatal Ikebukuro accident gets 5 years in prison | The Asahi Shimbun: Breaking News, Japan News and Analysis. In: The Asahi Shimbun [Internet]. [cited 28 Apr 2022]. Available: https://www.asahi.com/ajw/articles/14431946

35. Fournier T. Will My Next Car Be a Libertarian or a Utilitarian?: Who Will Decide? IEEE Technol Soc Mag. 2016;35: 40–45. https://doi.org/10.1109/MTS.2016.2554441

36. Savulescu J, Vergano M, Craxì L, Wilkinson D. An ethical algorithm for rationing life-sustaining treatment during the COVID-19 pandemic. BJA Br J Anaesth. 2020;125: 253–258. https://doi.org/10.1016/j.bja.2020.05.028 PMID:32560912

37. Contissa G, Lagioia F, Sartor G. The Ethical Knob: ethically-customisable automated vehicles and the law. Artif Intell Law. 2017;25: 365–378. https://doi.org/10.1007/s10506-017-9211-z

38. Christman J. Autonomy in Moral and Political Philosophy. Fall 2020. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy. Fall 2020. Metaphysics Research Lab, Stanford University; 2020. Available: https://plato.stanford.edu/archives/fall2020/entries/autonomy-moral/

39. Cabinet Secretariat. Social Principles of Human-Centric AI. 2020.

40. Arora C, Savulescu J, Maslen H, Wilkinson D. The Intensive Care Lifeboat: a survey of lay attitudes to rationing dilemmas in neonatal intensive care. BMC Med Ethics. 2016;17. https://doi.org/10.1186/s12910-016-0152-y PMID:27821118

41. Savulescu J, Gyngell C, Kahane G. Collective Reflective Equilibrium in Practice (CREP) and controversial novel technologies. Bioethics. 2021;35: 652–663. https://doi.org/10.1111/bioe.12869 PMID:33945162

8.10 Discussion Topics

The main theme for this chapter's discussion is about Professional Ethics & Responsibilities. Here are some suggested discussion points:

1. The Ethical Responsibilities of Professionals Working in the Technology Industry
   - What ethical obligations do technology professionals have toward users, employers, and society?
   - How should engineers balance innovation with user privacy and safety?
   - Are tech companies responsible for the unintended consequences of the products they release?
2. Examining Contrasting Ethical Perspectives
   - How do utilitarianism and deontology lead to different decisions in technology design?
   - Should AI systems be built with a specific ethical framework? If so, which one?
   - Can cultural differences justify conflicting ethical practices in global tech companies?
3. Examples of Ethical Violations and Their Impact on Society
   - What lessons can be learned from real-world cases like Therac-25, Facebook-Cambridge Analytica, or the VW emissions scandal?
   - What role did lack of accountability play in these violations?
   - How can we prevent future harm by understanding these past mistakes?
4. The Role of Professional Organizations in Promoting Ethical Behavior
   - Are codes of ethics (like those from ACM or IEEE) effective in shaping behavior in the tech industry?
   - How should professional organizations respond to violations by their members?
   - Should membership in a professional organization be mandatory to practice in tech?
5. Strategies for Maintaining Ethical Standards in the Technology Industry
   - What internal systems (e.g., ethical review boards, audits) are most effective in upholding ethics?
   - How can companies encourage employees to speak up about unethical practices?
   - How can tech teams embed ethics into the design and development process?
6. Whistleblowing: Exposing Wrongdoing vs. Honor in Professionalism
   - When is whistleblowing ethically justified, and what protections should whistleblowers have?
   - Is loyalty to a company ever more important than loyalty to the public good?
   - How do power dynamics influence the decision to blow the whistle?

Make use of the knowledge you gain from this book and your own experience, opinion and judgment will make your post more attractive for discussion.

8.11 Projects

8.11.1 Emotion Detection in Communication Devices: Applications, Implications, and Ethical Guidelines

Abstract: The project assesses the creation of an emotion detection platform for smartphones and other devices through evaluation of its uses and abuse potential. The study initiates its investigation through an overview of promising applications in different fields. The critical evaluation of potential risks together with negative consequences of emotion detection will be performed by the social scientists and privacy watchdogs. The combined team of technical experts and ethicists will develop guidelines for the ethical implementation of this technology by balancing innovation with privacy and ethical considerations.

Research Questions: An emotion detection system integrated with smartphones and other communication devices shows what advantages can be achieved in various fields. The study examines both the negative applications and privacy risks that emotion detection technology might present. A set of guidelines exists to protect privacy and consent during the ethical deployment of emotion detection systems.

Methodology: Application Research: The project identifies and describes various ways emotion detection technology can benefit healthcare, customer service, and personal well-being. The social scientists must perform a complete evaluation of privacy risks as well as the potential manipulative or abusive uses of the technology. The team will work with ethicists to build an ethical framework for emotion detection technology that includes consent protocols alongside transparency measures and data protection standards. The team should create policy recommendations to deploy the technology responsibly in society.

Expected Outcomes: This report provides detailed information about beneficial applications for emotion detection technology together with its advanced capabilities. The document examines both risks and ethical challenges that arise when emotion detection technology is misused. The project will establish strong privacy protection standards for ethical emotion detection system usage. The policy recommendations developed by this research can serve as guidelines for developers and lawmakers together with regulatory bodies to monitor this technology.

Timeline:

- Week 1: Project Planning and Potential Applications Research. A project outline must be created to define goals along with research methods and deliverable expectations. The team starts their research on beneficial applications of emotion detection technology.
- Week 2: Privacy Risks and Abuse Potential Assessment. The analysis focuses on privacy risks while exploring the negative potential uses of the technology. The team should start creating potential misuse scenarios along with their predicted societal consequences.
- Week 3: Ethical Guidelines Formulation. The researchers will work with ethicists to create guidelines for using emotion detection systems ethically. The researchers will combine findings from application research with risk assessment results while considering ethical standards.
- Week 4: Compilation and Policy Proposal Development. Finalize the set of ethical guidelines. The team needs to create a policy document which outlines the regulations for emotion detection technology use. The team will create a presentation to summarize the project's results alongside guidelines and policy recommendations.

Conclusion: The project aims to foster a comprehensive understanding of the multifaceted nature of emotion detection technology in communication devices, paving the way for its responsible development and application in society.

8.11.2 Curing the Common Cold: Ethics, Economics, and Employment

Abstract: This project considers the ethical dimensions as well as the financial implications of a hypothetical common cold cure discovery on workers within the cold-medicine industry. The research delves into whether the discoverer must conceal the cure to safeguard industrial employment and how society should understand the plight of workers whose jobs may disappear due to medical breakthroughs. This research will assess how public health advantages compare to economic consequences affecting people working in the sector.

Research Questions: When a cure for the common cold becomes available should it remain hidden to save the jobs of cold-medicine industry workers? What moral principles should guide the decision to share or hide major medical discoveries? What methods does society need to use in order to show compassion toward the multifaceted emotions workers experience when their jobs become obsolete due to public benefits of medical discoveries?

Methodology: **A** literature review of ethical principles will be performed to analyze medical discoveries and their effects on employment and economics.

An analysis of various stakeholders' viewpoints including medical personnel along with economists together with workers from the cold-medicine sector and the general public will be conducted.

The project will gather qualitative information about different groups' views through surveys and interviews to determine their attitudes and ethical positions on this matter.

Economic Impact Study: A review of economic models and historical precedents regarding medical breakthroughs and their impact on industries will be conducted.

Expected Outcomes: The research will deliver an exhaustive document which compares ethical implications of cold cure distribution to economic effects on the workforce. A compilation of stakeholder perspectives, providing insight into the varied reactions and emotions regarding industry disruption. The research will establish ethical recommendations which address both the overall public benefit and the well-being of workers affected by changes. The society should develop methods to show empathy toward workers who lose their jobs because of medical and technological developments.

Timeline:

- Week 1: Project Initiation and Literature Review. Develop project scope and methodology. Start the literature review to establish the ethical and economic background for the study.
- Week 2: Stakeholder Analysis and Data Collection Preparation. Determine which stakeholders need evaluation before creating stakeholder analysis procedures. Develop surveys and interview questions to determine stakeholder views regarding this topic.
- Week 3: Data Collection and Economic Impact Review. Perform surveys and interviews with stakeholders. The analysis of economic data will combine findings from previous medical breakthroughs.
- Week 4: Data Analysis and Report Writing. Combine ethical insights with economic results from collected data through analysis. The authors will create a final report which presents conclusions

alongside recommendations for action. A presentation will outline the research outcomes and introduce proposed methods to generate empathy along with support.

Conclusion: Students will examine the ethical conflict between medical innovation and economic security of people who work in current industries through this project. Students will learn to understand both the social implications of scientific progress and develop empathy toward its human effects.

8.12 Exercises

1. Why is data privacy a critical ethical responsibility for technology professionals?
2. What are the consequences of algorithmic bias, and how can professionals address them?
3. How can transparency and accountability be promoted in complex technological systems?
4. What role should technology professionals play in minimizing the social and environmental impacts of innovation?
5. How can professionals ethically respond to automation's impact on jobs?
6. How does Utilitarianism evaluate whether an action is ethical?
7. What is the main principle behind Deontological ethics according to Immanuel Kant?
8. What distinguishes Virtue Ethics from other ethical theories?
9. Why is Care Ethics considered a challenge to traditional ethical theories?
10. What is the ethical significance of freedom in Existentialist ethics?
11. What made the Enron scandal a significant ethical violation, and what were its consequences?
12. How did the Volkswagen emissions scandal affect public health and environmental standards?
13. What ethical concerns did the Cambridge Analytica scandal raise about personal data and democracy?
14. Why is the opioid crisis considered a failure of medical ethics?
15. What ethical dilemma does government surveillance pose for society?
16. How do professional organizations promote ethical behavior through codes of ethics?
17. What educational resources do professional organizations provide to support ethical decision-making?
18. In what ways do professional organizations foster accountability among members?
19. How do professional organizations influence public policy regarding ethical practices?
20. What role do professional organizations play in resolving ethical dilemmas?
21. Why are comprehensive ethical frameworks essential for the tech industry?
22. How does ongoing education contribute to ethical behavior in tech?
23. What is the role of transparency in ethical technology development?
24. How do ethical impact assessments benefit society?

25. Why is cross-disciplinary collaboration important for addressing tech ethics?
26. When is whistleblowing considered ethically justified?
27. How do professional ethics support whistleblowing?
28. What risks do whistleblowers often face?
29. How has the digital age influenced whistleblowing?
30. What are the broader impacts of whistleblowing on society?

Bibliography

[1] Acemoglu, D., Aghion, P., and Violante, G. L. *Deunionization, technical change and inequality, in Carnegie-Rochester conference series on public policy. 2001.* Elsevier.

[2] Acemoglu, D., and Autor, D. Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of labor economics, volume 4.* 2011, pp. 1043–1171.

[3] Acemoglu, D., and Restrepo, P. Demographics and automation. Working paper, NBER, 2012.

[4] Acemoglu, D., and Restrepo, P. Artificial intelligence, automation and work. Working paper, NBER, 2018.

[5] Acemoglu, D., and Restrepo, P. Robots and jobs: Evidence from us labor markets. *Journal of Political Economy 128,* 6 (2020), 2188–2244.

[6] Acemoglu, D., and Restrepo, P. Unpacking skill bias: Automation and new tasks, national bureau of economic research. 24. *Bessen, J., et al. (2019,* 90 (2020), 141–152.

[7] Acemoglu, D., and Tasks, R. P. automation, and the rise in us wage inequality. *NBER Work Pap [Internet] No. 2892* (2021).

[8] Aghion, P. et al., creative destruction and subjective well-being. *American Economic Review 106,* 12 (2016), 3869–97.

[9] Agrawal, A., Horton, J., Lacetera, N., and Lyons, E. Digitization and the contract labor market: A research agenda. In *Economic analysis of the digital economy, . University of Press.* 2015, pp. 219–250.

[10] Agrawal, A., Lacetera, N., and Lyons, E. Does standardized information in online markets disproportionately benefit job applicants from less developed countries? *Journal of international Economics 103* (2016), 1–12.

[11] Akerman, A., Gaarder, L, and Mogstad, M. The skill complementarity of broadband internet. *The Quarterly Journal of Economics 130,* 4 (2015), 1781–1824.

[12] Alharbi, H. A., Elgorashi, T. E., and Elmirghani, J. M. Impact of the net neutrality repeal on communication networks. *IEEE Access 8* (2020), 59787–59800.

[13] Allon, G., Cohen, M. C., and Sinchaisri, W. P. The impact of behavioral and economic drivers on gig economy workers. *Manufacturing & Service Operations Management* (2023).

[14] Alnajrani, H. M., Norman, A. A., and Ahmed, B. H. Privacy and data protection in mobile cloud computing: A systematic mapping study. *PLOS ONE 15, 6* (06 2020), 1–28.

[15] Aloisi, A., and Gramano, E. Artificial intelligence is watching you at work: Digital surveillance, employee monitoring, and regulatory issues in the eu context. *Comp. Lab. L. & Pol'y J. 41* (2019), 95.

[16] Alwyn, Y. Inequality, the urban-rural gap, and migration. *The Quarterly Journal of Economics 128,* 4 (2013), 1727–1785.

[17] Amanda, P., and Glassberg, S. E. Why the referential treatment? evidence from field experiments on referrals. *Journal of Political Economy 124,* 6 (2016), 1793–1828.

[18] Amir, A. M., and Mark, G. Hidden transcripts of the gig economy: labour agency and the new art of resistance among african gig workers. *Environment and Planning A: Economy and Space 52,* 7 (2020), 1269–1291.

[19] And, B. J. M. T. I. and income inequality: Socio-economic challenges in a hyperconnected society. *Telecommunications Policy 42,* 4 (2018), 333–343.

[20] Anderson, K. A. *Skill networks and measures of complex human capital.* Proceedings of the National Academy of Sciences, 2017.

[21] Angrist, J. D., and J.-S., P. *Mostly harmless econometrics: An empiricist's companion,* vol. Princeton. Princeton university press, NJ., 2009.

[22] Anner, M., Pons-Vignon, N., and Rani, U. For a future of work with dignity: A critique of the world 697 bank development report, the changing nature of work. *Global Labour Journal 10* (2019), 1.

[23] Arntz, M., Gregory, T., and Zierahn, U. Revisiting the risk of automation. *Econ LettOct 1,* 159 (2017), 157–60.

[24] Autor, D. H. Why are there still so many jobs? the history and future of workplace automation. *Journal of economic perspectives 29,* 3 (2015), 3–30.

[25] Autor, D. H., and Dorn, D. The growth of low-skill service jobs and the polarization of the us labor market. *The American Economic Review 103,* 5 (2013), 1553–1597.

[26] Autor, D. H., Dorn, D., Trade, H. G. U., and Evidence, T. from local labour markets. *Econ J 125,* 584 (2015), 621–46.

[27] Autor David, H., Frank, L., and Murnane Richard, J. The skill content of recent technological change: An empirical exploration. *The Quarterly journal of economics 118,* 4 (2003), 1279–1333.

[28] Ball, K., et al. Electronic monitoring and surveillance in the workplace. *European Commission Joint Research Centre* (2021).

[29] Barach Moshe, A., Golden Joseph, M., and Horton John, J. Steering in online markets: the role of platform incentives and credibility. *Management Science 66,* 9 (2020), 4047–4070.

[30] Barnidge, M., Kim, B., Sherrill, L. A., Luknar, Ž., and Zhang, J. Perceived exposure to and avoidance of hate speech in various communication settings. *Telematics and informatics 44* (2019), 101263.

[31] Barrero, J. M., Bloom, N., and Davis, S. J. Why working from home will stick. Technical report, National Bureau of Economic Research, 2021.

[32] Beaudry, P., Green, D. A., and Sand, B. M. The great reversal in the demand for skill and cognitive tasks. *Journal of Labor Economics 34* (2016), S199–S247.

[33] Becker, T. E. Potential problems in the statistical control of variables in organizational research: A qualitative analysis with recommendations. *Organizational Research Methods 8,* 3 (2005), 274–289.

[34] Beerepoot, N., and Lambregts, B. Competition in online job marketplaces: towards a global labour market for outsourcing services? *Global Networks 15,* 2 (2015), 236–255.

[35] Belfield, C. R., and Harris, R. D. How well do theories of job matching explain variations in job satisfaction across education levels? evidence for uk graduates. *Applied economics 34,* 5 (2002), 535–548.

[36] Benedek, W., and Kettemann, M. C. *Freedom of expression and the internet: Updated and revised 2nd edition.* Council of Europe, 2020.

[37] Benedikt, F. C., and Osborne Michael, A. The future of employment: how susceptible are jobs to computerisation? *Technological forecasting and social change 114* (2017), 254–280.

[38] Benson, A., Sojourner, A., and Umyarov, A. Can reputation discipline the gig economy? experimental evidence from an online labor market. *Management Science 66,* 5 (2020), 1802–1825.

[39] Bettencourt Luís, M. A., Jose, L., Deborah, S., and West Geoffrey, B. Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. *PloS one 5* (2010), 11.

[40] Bettencourt Luís, M. A., Jose, L., Dirk, H., Christian, K., and Growth, W. G. B. innovation, scaling, and the pace of life in cities. In *Proceedings of the national academy of sciences* (104(17), 2007), pp. 7301–7306.

[41] Bettencourt Luís, M. A., Samaniego, H., and Youn, H. Professional diversity and the productivity of cities. *Scientific reports 4,* 1 (2014), 1–6.

[42] Bettencourt Luís, M. A., Samaniego, H., and Youn, H. Professional diversity and the productivity of cities. *Scientific reports 4* (2014), 5393.

[43] Bhuller, M. et al., broadband internet: An information superhighway to sex crime? *Review of Economic Studies 80,* 4 (2013), 1237–1266.

[44] Blinder Alan, S. How many us jobs might be offshorable? *World Economics 10* (2009), 2.

[45] Blinder Alan, S., and Krueger Alan, B. Alternative measures of offshorability: a survey approach. *Journal of Labor Economics 31* (2013), S97–S128.

[46] B" Ockerman, P., Ilmakunnas, P., and Johansson, E. Job security and employee well-being: Evidence from matched survey and register data. *Labour Economics 18,* 4 (2011), 547–554.

[47] Boland, B., Smet, A. D., Palter, R., and Sanghvi, A. *Reimagining the office and work life after covid-19.* White paper, McKinsey, 2020.

[48] Borchert, K., Hirth, M., Kummer, M. E., Laitenberger, U., Slivko, O., and Viete, S. Unemployment and online labor. Tech, rep., ZEW-Centre for European Economic Research Discussion Paper, (18–023), 2018.

[49] Boschma Ron, A., and Michael, F. Creative class and regional growth: Empirical evidence from seven european countries. *Economic geography 85,* 4 (2009), 391–423.

[50] Braesemann, F., Lehdonvirta, V., and K" Assi, O. lets and the urban-rural divide: can online labour platforms bridge the gap? *Information, Communication & Society 25,* 1 (2022), 34–54.

[51] Braesemann, F., Stephany, F., Teutloff, O., Kã€ssi, O., Graham, M., and Lehdonvirta, V. The global polarisation of remote work. *PLOS ONE 17,* 10 (10 2022), 1–22.

[52] Braunack-Mayer, A., Carolan, L., Street, J., Ha, T., Fabrianesi, B., and Carter, S. Ethical issues in big data: A qualitative study comparing responses in the health and higher education sectors. *PLOS ONE 18,* 4 (04 2023), 1–16.

[53] Brown, P., Lauder, H., and Ashton, D. *The global auction: The broken promises of education, jobs, and incomes.* Oxford University Press, 2010.

[54] Brynjolfsson, E., Mitchell, T., and Rock, D. What can machines learn and what does it mean for occupations and the economy? *Am Econ Assoc 108* (2018), 43–7.

[55] Burbidge John, B., Lonnie, M., and Leslie, R. A. Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association 83,* 401 (1988), 123–127.

[56] Cameron, C. A., and Miller D. L., A. practitioner's guide to cluster-robust inference. *Journal of Human Resources 50,* 2 (2015), 317–372.

[57] Campbell-Kelly, M., Aspray, W., Yost, J., Tinn, H., and Díaz, G. *Computer: A History of the Information Machine.* Taylor & Francis, 2023.

[58] Candeub, A. Bargaining for free speech: Common carriage, network neutrality, and section 230. *Yale JL & Tech. 22* (2020), 391.

[59] Carnevale Joel, B., and Isabella, H. Employee adjustment and well-being in the era of covid-19: Implications for human resource management. *Journal of Business Research 116* (2020), 183–187.

[60] Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., and López, H. M. H. Internet, social media and online hate speech, systematic review. *Aggression and Violent Behavior 58* (2021), 101608.

[61] Ceruzzi, P. E. *A history of modern computing.* MIT press, 2003.

[62] Chiacchio, F., Petropoulos, G., and Pichler, D. The impact of industrial robots on eu employment and wages: A local labour market approach. Tech, rep., Bruegel Working Papers (02), 2018.

[63] Cho, J.-C. Human microbiome privacy risks associated with summary statistics. *PLOS ONE 16,* 4 (04 2021), 1–11.

[64] Chris, F., and Van Zeebroeck Nicolas. From wires to partners: How the internet has fostered r&d collaborations within firms. *Management science 58,* 8 (2012), 1549–1568.

[65] Christina, B., Anh, N. Q., and Uma, R. Social protection systems and the future of work: Ensuring social security for digital platform workers. *International Social Security Review 72,* 3 (2019), 17–41.

[66] Clark, A., Oswald, A., and Warr, P. Is job satisfaction ushaped in age? *Journal of occupational and organizational psychology 69,* 1 (1996), 57–81.

[67] Cohen, A. Could better technology lead to stronger 4th amendment privacy protections? Tech, rep., Brennan Center, 2022.

[68] Collier, P. The future of capitalism: Facing the new anxieties penguin uk.

[69] Commission, E. Digital Single Market Available from, 2014.

[70] CORRADO, C., and Van Ark, B. *The internet and productivity.* Edward Elgar, In Handbook on the Economics of the Internet. Publishing, 2016.

[71] CORTEZ, E. *Data Protection Around the World: Privacy Laws in Action.* Information Technology and Law Series. T.M.C. Asser Press, 2020.

[72] Danezis, G., Domingo-Ferrer, J., Hansen, M., Hoepman, J.-H., Metayer, D. L., Tirtea, R., and Schiffner, S. Privacy and data protection by design-from policy to engineering. *arXiv preprint arXiv:1501.03726* (2015).

[73] Dauth, W. et al., adjusting to robots: Worker-level evidence. Working paper, Opportunity and Inclusive Growth Institute, 2014.

[74] David, H., and Dorn, D. The growth of low-skill service jobs and the polarization of the us labor market. *American Economic Review 103,* 5 (2013), 1553–97.

[75] David, H., Katz Lawrence, F., and Kearney Melissa, S. The polarization of the us labor market. *American economic review 96,* 2 (2006), 189–194.

[76] David, S., and Ivana, P. Algorithmic management in the platform economy. *Sociologica 14,* 3 (2020), 47–72.

[77] De Gregorio, G. Democratising online content moderation: A constitutional framework. *Computer Law & Security Review 36* (2020), 105374.

[78] Deming, D., and LB., K. Skill requirements across firms and labor markets: Evidence from job postings for professionals. *J Labor Econ 36* (2018).

[79] Diodato, D., Neffke, F., and O'Clery, N. Why do industries coagglomerate? how marshallian externalities differ by industry and have evolved over time. *Journal of Urban Economics 106* (2018), 1–26.

[80] Diversity and, Q. J. M. U. and economic growth. *Journal of Economic perspectives 12,* 2 (1998), 127–138.

[81] DJ., D. The growing importance of digital skills in the workplace. *Q J Econ 132* (October 2017), 1593–640.

[82] Djumalieva, J., Lima, A., and Sleeman, C. Classifying occupations according to their skill requirements in job advertisements. *ESCoE Discuss Pap [Internet]. 2018* (Jul 2021).

[83] Douek, E. Governing online speech: From "posts-as-trumps" to proportionality and probability. *Colum. L. Rev. 121* (2021), 759.

[84] Ehrnsperger, J. F., and Tietze, F. Patent pledges, open ip, or patent pools? developing taxonomies in the thicket of terminologies. *PLOS ONE 14,* 8 (08 2019), 1–18.

[85] Ejaz, G., Kerr William, R., and Christopher, S. Diasporas and outsourcing: evidence from odesk and india. *Management Science 60,* 7 (2014), 1677–1697.

[86] Elkin-Koren, N., and Perel, M. Separation of functions for ai: Restraining speech regulation by online platforms. *Lewis & Clark L. Rev. 24* (2020), 857.

[87] Erdogan, B. et al., whistle while you work: A review of the life satisfaction literature. *Journal of management 38,* 4 (2012), 1038–1083.

[88] Fabian, S., Michael, D., Steven, S., and Vili, L. Distancing bonus or downscaling loss? the changing livelihood of us online workers in times of covid-19. *Tijdschrift voor economische en sociale geografie 111,* 3 (2020), 561–573.

[89] Facca, D., Smith, M. J., Shelley, J., Lizotte, D., and Donelle, L. Exploring the ethical issues in research using digital data collection strategies with minors: A scoping review. *PLOS ONE 15,* 8 (08 2020), 1–17.

[90] Falk, M., and Biagi, F. Relative demand for highly skilled workers and use of different ict technologies. *Applied Economics 49,* 9 (2017), 903–914.

[91] Fast, A. A., Olson, K. R., and Mandel, G. N. Intuitive intellectual property law: A nationally-representative test of the plagiarism fallacy. *PLOS ONE 12,* 9 (09 2017), 1–11.

[92] Feakins, M. Off-offshoring from russia to Ukraine: How russian transnational entrepreneurs created a post-soviet it offshore. *Economic Geography 93,* 4 (2017), 397–417.

[93] Fernández-Macías, E., Hurley, J., and Bisello, M. What do europeans do at work?: A task-based analysis: European jobs monitor 2016. *Publ Off Eur Union [Internet], 2016* (Jul 2021).

[94] Flonk, D., Jachtenfuchs, M., and Obendiek, A. S. Authority conflicts in internet governance: Liberals vs. sovereigntists? *Global Constitutionalism 9,* 2 (2020), 364–386.

[95] Florida, R. L., et al. *Cities and the creative class.* Psychology Press, 2005.

[96] Forman, C., Goldfarb, A., and Greenstein, S. How geography shapes—And is shapey by—the internet. In *The New Handbook of Economic Geography.* Oxford University Press, Oxford, 2018, pp. 269–285.

[97] Frank, M. R. et al., small cities face greater impact from automation. *Journal of the Royal Society Interface 15* (2018), 139.

[98] Frank, M. R. et al., toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences 116,* 14 (2019), 6531–6539.

[99] Frey, C. B., and MA., O. The future of employment: how susceptible are jobs to computerization? *Technol Forecast Soc Change* (2013), 1–72.

[100] Frey, C. B., and MA., O. The future of employment: How susceptible are jobs to computerisation? *Technol Forecast Soc ChangeJan 1,* 114 (2017), 254–80.

[101] Gallardo, R., Whitacre, B., Kumar, I., and Upendram, S. Broadband metrics and job productivity: a look at county-level data. *The Annals of Regional Science 66,* 1 (2021), 161–184.

[102] George, J. F., Chen, R., and Yuan, L. Intent to purchase iot home security devices: Fear vs privacy. *PLOS ONE 16,* 9 (09 2021), 1–14.

[103] Goos, M., Arntz, M., Zierahn, U., Gregory, T., Carretero Gómez, S., González Vázquez, I., et al. The impact of technological innovation on the future of work. *JRC Work Pap Ser Labour, Educ Technol [Internet], 2019* (Jul 2021).

[104] Goos, M., Manning, A., and Salomons, A. Explaining job polarization: Routine-biased technological change and offshoring. *American Economic Review 10],* 8 (2014), 2509–26.

[105] Goos, M., Manning, A., and Salomons, A. Explaining job polarization: Routine-biased technological change and offshoring. *American economic review 10],* 8 (2014), 2509–26.

[106] Gorwa, R. The platform governance triangle: Conceptualising the informal regulation of online content. *Internet Policy Review 8,* 2 (2019), 1–22.

[107] Graetz, G., and Michaels, G. Robots at work. *The Review of Economics and Statistics 100,* 5 (2018), 753–768.

[108] Graham, M. Warped geographies of development: The internet and theories of economic development. *Geography compass 2,* 3 (2008), 771–789.

[109] Graham, M., Hjorth, I., and Lehdonvirta, V. Digital labour and development: impacts of global digital labour platforms and the gig economy on worker livelihoods. *Transfer: European review of labour and research 23,* 2 (2017), 135–162.

[110] Green, F., and Zhu, Y. Overqualification, job dissatisfaction and increasing dispersion in the returns to graduate education. *Oxford economic papers 62,* 4 (2010), 740–763.

[111] Gregory, T., Salomons, A., and Zierahn, U. *Racing With or Against the Machine?* Evidence from Europe. IZA Discuss Pap, 2019.

[112] Han, F., Zhang, S., Yuan, J., and Wang, L. Assessing future technological impacts of patents based on the classification algorithms in machine learning: The case of electric vehicle domain. *PLOS ONE 17,* 12 (12 2022), 1–17.

[113] Harrington, D. U.S. privacy laws: The complete guide. Working paper, Varonis, 2023.

[114] Hecht-Felella, L. The fourth amendment in the digital age how carpenter can shape privacy protections for new technologies. Tech, rep., Brennan Center for Justice at New York University School of Law, 2021.

[115] Heckman, J., Fostering, K. T., and Skills, M. *Interventions that improve character and cognition.* Nber Work Pap Ser, 2013.

[116] Heeks, R. Decent work and the digital gig economy. Working Paper 71, Development Informatics, 2017.

[117] Helberger, N. The political power of platforms: How current attempts to regulate misinformation amplify opinion power. *Digital Journalism 8,* 6 (2020), 842–854.

[118] Horton, J., Kerr, W. R., and Stanton, C. High-skilled migration to the united states and its economic consequences. In *editors, High-Skilled Migration to the United States and Its Economic Consequences, chapter 3,* G. H. Hanson, W. R. Kerr, and S. Turner, Eds. University of Press, Chicago, 2018, pp. 71–108.

[119] Horton John, J. Buyer uncertainty about seller capacity: Causes, consequences, and a partial solution. *Management Science 65,* 8 (2019), 3518–3540.

[120] Houseman, S. N. Understanding the decline of us manufacturing employment. W.E. Upjohn Institute for Employment Research. Upjohn Institute working paper, 2018.

[121] Hunter, J. E., Schmidt, F. L., and Judiesch, M. K. Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology 75,* 1 (1990), 28–42.

[122] Inequality and, G. A. K. I. and growing disparity: Spatial patterns of inequality and the case of the usa. *The new Oxford handbook of economic geography, page 63* (2018).

[123] Innovation Vivarelli, M. Employment and skills in advanced and developing countries: A survey of economic literature. *J Econ Issues [Internet], 48,* 1 (Nov 2019), 123–54.

[124] Ipeirotis, P. G., and Horton, J. J. *The need for standardization in crowdsourcing.* In Proceedings of the workshop on crowdsourcing and human computation at CHI, 2011.

[125] Janssen, C. P., Donker, S. F., Brumby, D. P., and Kun, A. L. History and future of human-automation interaction. *International journal of human-computer studies 131* (2019), 99–107.

[126] Jiang, H., Tsohou, A., Siponen, M., and Li, Y. Examining the side effects of organizational internet monitoring on employees. *Internet Research 30,* 6 (2020), 1613–1630.

[127] Jiang, J. A., Scheuerman, M. K., Fiesler, C., and Brubaker, J. R. Understanding international perceptions of the severity of harmful content online. *PLOS ONE 16, 8* (08 2021), 1–22.

[128] Jonas, H., and Jonas, P. The arrival of fast internet and employment in africa. *American Economic Review 109,* 3 (2019), 1032–79.

[129] Josten, C., At Work: Automatable, L. G. R., and Automatable Jobs, N. In: Handbook of labor, human resources and population economics. In *International Publishing; p.* 2020, pp. 1–24.

[130] JOSTEN, C., and Lordan, G. Automation and the changing nature of work. *PLOS ONE 17, 5* (05 2022), 1–15.

[131] Kahneman, D., and Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica 47,* 2 (1979), 263–291.

[132] Kaine, S., and Josserand, E. The organisation and experience of work in the gig economy. *Journal of Industrial Relations 61,* 4 (2019), 479–501.

[133] K" ASSI, O., LEHDONVIRTA, V., and Stephany, F. *How many online workers are there in the world? a data-driven assessment [version 2* (2021).

[134] Kautz, T., Heckman, J. J., Diris, R., Ter, W. B., Fostering, B. L. ., and Skills, M. Improving cognitive and non-cognitive skills to promote lifetime success. *Natl Bur Econ Res [Internet], 2014* (Dec 2021), 15.

[135] Khuong, V., Payam, H., and Erik, B. let as a driver of economic growth: A survey of the literature and directions for future research. *Telecommunications Policy 44* (2020), 2.

[136] Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., et al. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work* (2013), pp. 1301–1318.

[137] Koops, B.-J., Newell, B. C., Timan, T., Skorvanek, I., Chokrevski, T., and Galic, M. A typology of privacy. *U. Pa. J. Int'l L. 38* (2016), 483.

[138] Kotschi, J., and Horneburg, B. The open source seed licence: A novel approach to safeguarding access to plant germplasm. *PLOS Biology 16,* 10 (10 2018), 1–7.

[139] Lamo, M., and Calo, R. Regulating bot speech. *UCLA L. Rev. 66* (2019), 988.

[140] Lehdonvirta, V., Barnard, H., Graham, M., and Hjorth, I. *Online labour markets-levelling the playing field for international service markets?* Conference paper, Oxford Internet Institute, 2014.

[141] Lehdonvirta, V., K" Assi, O., Hjorth, L, Barnard, H., and Graham, M. The global platform economy: A new offshoring institution enabling emerging-economy microproviders. *Journal of Management 45,* 2 (2019), 567–599.

[142] Leo, S. The productivity of cities. *The Quarterly Journal of Economics 89,* 3 (1975), 393–413.

[143] Levis, D., Fontana, F., and Ughetto, E. A look into the future of blockchain technology. *PLOS ONE 16,* 11 (11 2021), 1–20.

[144] Lewandowski, P., Park, A., Hardy, W., and Technology, Y. D. Skills, and globalization: Explaining international differences in routine and nonroutine work using survey data. *IZA Discuss Pap [Internet], 2019* (Jul 2021).

[145] Lordan, G. *Robots at work—A report on automatable and non-automatable employment shares in Europe.* Eur Comm—Dir Employment, Soc Aff Incl, 2018.

[146] Lordan, G., and J-S., P. Does rosie like riveting? male and female occupational choices. *NBER Work Pap [Internet], 2016* (Nov 2019).

[147] Lordan, G., and Mcguire, A. Widening the high school curriculum to include soft skill training: Impacts on health, behaviour, emotional wellbeing and occupational aspirations. *IZA Discuss Pap (1243* (2019), 9.

[148] Lordan, G., and Versus Machines, N. D. P. The impact of minimum wages on automatable jobs. *Labour EconJun 1,* 52 (2018), 40–53.

[149] Lowenberg, D., and Puebla, I. Responsible handling of ethics in data publication. *PLOS Biology 20,* 3 (03 2022), 1–3.

[150] Lukac, M. Two worlds of online labour markets: Exploring segmentation using finite mixture models and a network of skill co-occurrence. Working paper, SocArXiv, 2021.

[151] Lupu, Y., Sear, R., VELÃjSQUEZ, N., Leahy, R., Restrepo, N. J., Goldberg, B., and Johnson, N. F. Offline events and online hate. *PLOS ONE 18,* 1 (01 2023), 1–14.

[152] MA., B. L. The origins of scaling in cities, science, 340(6139):1438–1441.

[153] Maddala, G. S., and Lee, L.-F. Recursive models with qualitative endogenous variables, *in annals of Economic and Social Measurement 5,* 4 (1976), 525–545.

[154] Mai, K. T., Bray, S., Davies, T., and Griffin, L. D. Warning: Humans cannot reliably detect speech deepfakes. *PLOS ONE 18,* 8 (08 2023), 1–20.

[155] Malecki Edward, J. The economic geography of the internet's infrastructure. *Economic geography 78,* 4 (2002), 399–424.

[156] Mallett Richard, W. Seeing the "changing nature of work" through a precarity len. *Global Labour Journal 11* (2020), 3.

[157] Manning, S., Larsen, M. M., and Kannothra, C. G. Global sourcing of business processes: History, effects, and future trends. In *The new handbook of economic geography, Publisher: Oxford.* Oxford University Press, Oxford, 2017, pp. 407–435.

[158] Manyika, J., Cabral, A., Moodley, L., Moraje, S., Yeboah-Amankwah, S., Chui, M., et al. Lions go digital: The internet's transformative potential in africa. Technical report, McKinsey Global Institute, 2013.

[159] Manyika, J., Lund, S., Robinson, K., Valentino, J., and Dobbs, R. A labor market that works: Connecting talent with opportunity in the digital age. Working paper, McKinsey Global Institute, 2015.

[160] Mark, G., and Amir, A. M. The global gig economy: Towards a planetary labour market? *First Monday 24* (2019), 4.

[161] Martin, L., and andre, G. Reputation systems and recruitment in online labor markets: insights from an agent-based model. *Journal of Computational Social Science 4,* 1 (2021), 207–229.

[162] Mcguinness, S., Pouliakas, K., and Redmond, P. Skills-displacing technological change and its impact on jobs: Challenging technological alarmism? iza discussion paper no. 12541. Working paper, 2019. 19. Klenert, D., E. Fernandez-Macias and J.-I. Anton. (2020). Do robots really destroy jobs? Evidence from Europe. Seville: European Commission. Joint Research Centre. 20. Blanas S., Gancia G. and Lee S.Y.T.. Who is afraid of machines? Barcelone GSE, 2019.

[163] Mihaylov, E., and Tijdens, K. G. Measuring the routine and non-routine task content of 427 four-digit isco-08 occupations. Working paper, SSRN, 2019.

[164] Mindell, D. A., Reynolds, E., et al. *The work of the future: building better jobs in an age of intelligent machines.* MIT Press, 2023.

[165] Mingfeng, L., Yong, L., and Siva, V. Effectiveness of reputation in contracting for customized production: Evidence from online labor markets. *Management Science 64,* 1 (2018), 345–359.

[166] Mohlmann, M., Zalmanson, L., Henfridsson, O., and Wayne, G. R. Algorithmic management of work on online labor platforms: When matching meets control. *MIS Quarterly 45* (2021), 4.

[167] Mokyr, J., Vickers, C., and Ziebarth, N. L. The history of technological anxiety and the future of economic growth: Is this time different? *Journal of Economic Perspectives 29,* 3 (2015), 31–50.

[168] Monfardini, C., and Radice, R. Testing exogeneity in the bivariate probit model: A monte carlo study. *Oxford Bulletin of Economics and Statistics 70,* 2 (2008), 271–282.

[169] Moulton, B. R. Random group effects and the precision of regression estimates. *Journal of Econometrics 32,* 3 (1986), 385–397.

[170] Mozafari, M., Farahbakhsh, R., and Crespi, N. Hate speech detection and racial bias mitigation in social media based on bert model. *PLOS ONE 15,* 8 (08 2020), 1–26.

[171] Muir, S. R., Roberts, L. D., Sheridan, L., and Coleman, A. R. Examining the role of moral, emotional, behavioural, and personality factors in predicting online shaming. *PLOS ONE 18,* 3 (03 2023), 1–19.

[172] Mulligan, S. P., Freeman, W. C., and Linebaugh, C. D. Data protection and privacy law: An introduction. Working paper, Congressional Research Service, 2022.

[173] Myers, B. A. A brief history of human-computer interaction technology, *interactions 5,* 2 (1998), 44–54.

[174] Nedelkoska, L., and Automation, Q. G. skills use and training (oecd. Tech, rep., Working Paper). OECD Soc Employment, Migr Work Pap [Internet] (202):1–125. Available from:, 2018.

[175] Oboyle, E. H., and Forsyth, D. R. Individual differences in ethics positions: The epq-5. *PLOS ONE 16,* 6 (06 2021), 1–18.

[176] Ocasio, M. A., Fernandez, M. I., Cortese, S., and Kampa, K. Piloting a digital campaign to promote awareness of the louisiana teleprep program among sexual and gender minority young adults. *PLOS ONE 18,* 8 (08 2023), 1–15.

[177] Ochang, P., Stahl, B. C., and Eke, D. The ethical and legal landscape of brain data governance. *PLOS ONE 17,* 12 (12 2022), 1–29.

[178] Okumoto, A., Yoneyama, S., Miyata, C., and Kinoshita, A. The relationship between hospital ethical climate and continuing education in nursing ethics. *PLOS ONE 17, 7* (07 2022), 1–15.

[179] Ostermaier, A., and Uhl, M. Spot on for liars! how public scrutiny influences ethical behavior. *PLOS ONE 12, 7* (07 2017), 1–11.

[180] Oswald, A. J., Proto, E., and Sgroi, D. Happiness and productivity. *Journal of Labor Economics 33,* 4 (2015), 789–822.

[181] Otto, K., and Vili, L. Online labour index: Measuring the online gig economy for policy and research. *Technological forecasting and social change 137* (2018), 241–248.

[182] Otto, K., and Vili, L. Do microcredentials help new workers enter the market? evidence from an online labor platform. *Journal of Human Resources, pages 03* (2022), 519–10226.

[183] Ozimek, A. The future of remote work. *Available at SSRN 3638597* (2020).

[184] Pajarinen, M., Rouvinen, P., Claussen, J., Hakanen, J., Kovalainen, A., Kretschmer, T., et al. Upworkers in finland: Survey results. Technical report, ETLA Report, 2018.

[185] PALLAIS, A. Inefficient hiring in entry-level labor markets. *American Economic Review 104,* H (2014), 3565–99.

[186] Perrons, D. Understanding social and spatial divisions in the new economy: New media clusters and the digital divide. *Economic geography 80,* 1 (2004), 45–61.

[187] Phiromswad, P., Srivannaboon, S., and Sarajoti, P. The interaction effects of automation and population aging on labor market. *PLOS ONE 17, 2* (02 2022), 1–16.

[188] Pickard, V., and Berman, D. E. *After net neutrality.* Yale University Press, 2019.

[189] Pierre-Alexandre, B., Jara-Figueroa, C., Petralia Sergio, G., Steijn Mathieu, P. A., Rigby David, L., and Hidalgo, C. A. Complex economic activities concentrate in large cities. *Nature Human Behaviour 4,* 3 (2020), 248–254.

[190] Reglitz, M. The human right to free internet access. *Journal of Applied Philosophy 37,* 2 (2020), 314–331.

[191] Richard, G. I., and Ioannis, K. Accounting for big-city growth in low-paid occupations: Immigration and/or service-class consumption. *Economic Geography 90,* 1 (2014), 67–90.

[192] Roodman, D. Fitting fully observed recursive mixed-process models with cmp. *Stata Journal 11,* 2 (2011), 159–206.

[193] Roy, G., and Shrivastava, A. K. Future of gig economy: opportunities and challenges. *IMI Konnect 9,* 1 (2020), 14–27.

[194] Sajaia, Z. Maximum likelihood estimation of a bivariate ordered probit model: implementation and monte carlo simulations. *The Stata Journal 4,* 2 (2008), 1–18.

[195] Salvatori, A. Labour contract regulations and workers' wellbeing: International longitudinal evidence. *Labour Economics 17, I* (2010), 667–678.

[196] Sander, B. Freedom of expression in the age of online platforms: The promise and pitfalls of a human rights-based approach to content moderation. *Fordham Int'l LJ 43* (2019), 939.

[197] Schwabe, H., and Castellacci, F. Automation, workersâ skills and job satisfaction. *PLOS ONE 15,* 11 (11 2020), 1–26.

[198] Seastedt, K. P., Schwab, P., Oâbrien, Z., Wakida, E., Herrera, K., Marcelo, P. G. F., Agha-Mir-Salim, L., Frigola, X. B., Ndulue, E. B., Marcelo, A., and Celi, L. A. Global healthcare fairness: We should be sharing more, not less, data. *PLOS Digital Health 1,* 10 (10 2022), 1–13.

[199] Seering, J. Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction 4,* CSCW2 (2020), 1–28.

[200] Skatova, A., Mcdonald, R., Ma, S., and Maple, C. Unpacking privacy: Valuation of personal data protection. *PLOS ONE 18, 5* (05 2023), 1–21.

[201] Skliaustyte, E., and Weber, M. Subsidies versus intellectual property rights when innovators operate in two markets. *PLOS ONE 18,* 4 (04 2023), 1–20.

[202] Smalls, D., and Wilson, G. Ten quick tips for staying safe online. *PLOS Computational Biology 17, 3* (03 2021), 1–6.

[203] Smits, J. Gdl area database. Working Paper 16, GDL, 2016.

[204] Soares, F. B., Gruzd, A., Jacobson, J., and Hodson, J. To troll or not to troll: Young adultsa anti-social behaviour on social media. *PLOS ONE 18,* 5 (05 2023), 1–17.

[205] Sousa-Poza, A., and Sousa-Poza, A. A. Well-being at work: a cross-national analysis of the levels and determinants of job satisfaction. *The journal of socio-economics 29,* 6 (2000), 517–538.

[206] Spector, P. E., and Brannick, M. T. Methodological urban legends: The misuse of statistical control variables. *Organizational Research Methods 14,* 2 (2011), 287–305.

[207] Stanton Christopher, T., and Catherine, T. Landing the first job: The value of intermediaries in online hiring. *The Review of Economic Studies 83,* 2 (2016), 810–854.

[208] Stephan, M., Larsen Marcus, M., and Kannothra, C. G. Global sourcing of business processes. In *The New Handbook of Economic Geography.* Oxford University Press, Oxford, 2018, pp. 407–426.

[209] Stjernfelt, F., and Lauritzen, A. M. *Your post has been removed: Tech giants and freedom of speech.* Springer Nature, 2020.

[210] Suominen, K. *Fuelling Trade in the Digital Era.* Issue Paper, International Centre for Trade and Sustainable Development, Geneva, 2017.

[211] Swierenga, R. P. Computers and american history: the impact of the“ new” generation. *The Journal of American History 60,* 4 (1974), 1045–1070.

[212] Takaguchi, K., Kappes, A., Yearsley, J. M., Sawai, T., Wilkinson, D. J. C., and Savulescu, J. Personal ethical settings for driverless cars and the utility paradox: An ethical analysis of public attitudes in uk and japan. *PLOS ONE 17,* 11 (11 2022), 1–19.

[213] Thiago, D. O., Marcelo, A. D., and Gomes, A. Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to Igbtq voices online. *Sexuality & culture 25,* 2 (2021), 700–732.

[214] Tolan, S., Pesole, A., Martínez-Plumed, F., Fernández-Macías, E., Hernández-Orallo, J., and Gómez, E. Measuring the occupational impact of ai: Tasks, cognitive abilities and ai benchmarks. *J Artif Intell Res 71* (2021), 191–236.

[215] Torosyan, K., Wang, S., Mack, E. A., Van Fossen, J. A., and Baker, N. Assessing the impact of technological change on similar occupations: Implications for employment alternatives. *PLOS ONE 18, 9* (09 2023), 1–27.

[216] Trivedi, S., and Patel, N. Virtual employee monitoring: A review on tools, opportunities, challenges, and decision factors. *Empirical Quests for Management Essences 1,* 1 (2021), 86–99.

[217] Tschang, F. T., and Almirall, E. Artificial intelligence as augmenting automation: Implications for employment. *Academy of Management Perspectives 35,* 4 (2021), 642–659.

[218] Uma, R., and Marianne, F. On-demand digital economy: Can experience ensure work and income security for microtask workers? *Jahrb “ucher fur Nationalokonomie und Statistik 239,* 3 (2019), 565–597.

[219] Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W., and Hoste, V. Automatic detection of cyberbullying in social media text. *PLOS ONE 13,* 10 (10 2018), 1–22.

[220] VlLl, L., and Greetje, C. *Platform Sourcing: How Fortune 500 Firms are Adopting Online Freelancing Platforms.* Internet Institute, Oxford, 2017.

[221] Waqas, A., Salminen, J., Jung, S.-G., Almerekhi, H., and Jansen, B. J. Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PLOS ONE 14,* 9 (09 2019), 1–21.

[222] Webb, M. *The Impact of Artificial Intelligence on the Labor Market.* SSRN, 2019.

[223] Weems, C. F., Ahmed, I., Richard, III, G. G., Russell, J. D., and Neill, E. L. Susceptibility and resilience to cyber threat: Findings from a scenario decision program to measure secure and insecure computing behavior. *PLOS ONE 13,* 12 (12 2018), 1–18.

[224] Weinberger, C. The increasing complementarity between cognitive and social skills. *Rev Econ Stat [Internet], 96* (Jan 2021), 5.

[225] Werner, E., Holger, H., Ulf, R., and Verena, T. How big is the gig? assessing the preliminary evidence on the effects of digitalization on the labor market. *Management revue* (2017), 298–318.

[226] Wilkowska, W., Offermann, J., Spinsante, S., Poli, A., and Ziefle, M. Analyzing technology acceptance and perception of privacy in ambient assisted living for using sensor-based technologies. *PLOS ONE 17,* 7 (07 2022), 1–21.

[227] Willcocks, L. Robo-apocalypse cancelled? reframing the automation and future of work debate. *Journal of Information Technology 35,* 4 (2020), 286–302.

[228] Wood Alex, J., Mark, G., Vili, L., and Isis, H. Good gig, bad gig: Autonomy and algorithmic control in the global gig economy. *Work, Employment and Society 33,* 1 (2019), 56–75.

[229] Wooldridge, J. M. Econometric analysis of cross section and panel data MIT press, 2010.

[230] Yeung, H. W.-C. The logic of production networks. In *The New Handbook of Economic Geography.* Oxford University Press, Oxford, 2018, pp. 382–406.

[231] Yili, H., and Pavlou Paul, A. On buyer selection of service providers in online outsourcing platforms for it services. *Information Systems Research 28,* 3 (2017), 547–562.

[232] Yoo, C. S. The first amendment, common carriers, and public accommodations: net neutrality, digital platforms, and privacy. *J. Free Speech L. 1* (2021), 463.

[233] Yoo, J., and Phillips, J. C. Technology and the fourth amendment. *National Review* (March 2019).

[234] Yuen, T. B. The digital nomad lifestyle:(remote) work/leisure balance, privilege, and constructed 661 community. *International Journal of the Sociology of Leisure 2,* 1 (2019), 27–42.

[235] Zannettou, S., Elsherief, M., Belding, E., Nilizadeh, S., and Stringhini, G. Measuring and characterizing hate speech on news websites. In *Proceedings of the 12th ACM Conference on Web Science* (2020), pp. 125–134.